Methodische und didaktische Überlegungen sowie empirische Befunde zur Erfassung sprachlicher Kompetenzen im Deutschen

Analysen zu den Bildungsstandards im Fach Deutsch für den Primarbereich

Dissertation

zur Erlangung des akademischen Grads

Dr. Phil.

im Fach Erziehungswissenschaften

eingereicht am 31.10.2011

verteidigt am 16.12.2011

an der Philosophischen Fakultät IV der Humboldt-Universität zu Berlin

von

Dipl.-Psych. Katrin Böhme

Präsident der Humboldt-Universität zu Berlin

Prof. Dr. Jan-Hendrik Olbertz

Dekan der Philosophischen Fakultät IV

Prof. Dr. Ernst von Kardorff

Gutachter:

1. Prof. Dr. Michael Kämper-van den Boogaart

2. Prof. Dr. Olaf Köller

Zusammenfassung

Die vorliegende Dissertation ist in der empirischen Bildungsforschung angesiedelt und beschäftigt sich mit der Erfassung sprachlicher Kompetenzen im Deutschen. Hierbei beinhaltet die Arbeit sowohl methodische als auch didaktische Überlegungen sowie die Präsentation empirischer Befunde, die sich auf Analysen zu den Bildungsstandards im Fach Deutsch für den Primarbereich beziehen. Die Dissertation umfasst vier empirische Beiträge, die von einer integrierenden Rahmung umschlossen werden.

Zunächst erfolgt eine Verortung der vorliegenden Arbeit in den wissenschaftlichen Referenzdisziplinen, eine Klärung des zentralen Kompetenzbegriffs, eine Beschäftigung mit messmethodischen Aspekten sowie eine vertiefte Auseinandersetzung mit den in den Bildungsstandards eingeführten sprachlichen Kompetenzen.

Der erste empirische Beitrag mit dem Titel „Zur Abgrenzung des Hörverstehens gegenüber dem Leseverstehen mit Hilfe schwierigkeitsbestimmender Merkmale bei der Entwicklung von Testaufgaben“ (Böhme, Robitzsch & Busè, 2010) thematisiert die Operationalisierung des Hörverstehens im Rahmen von Large-Scale-Assessments und stellt Ergebnisse zur Analyse schwierigkeitsbestimmender Merkmale in bildungsstandardbasierten Testaufgaben vor.

Der zweite empirische Beitrag trägt den Titel „Aspekte der Kodierung von Schreibaufgaben“ (Böhme, Bremerich-Vos & Robitzsch, 2009) und widmet sich diagnostischen Fragen der Operationalisierung der Schreibkompetenz und der Bewertung von Schreibprodukten. Neben der Untersuchung der Reliabilität der durch Rater vorgenommenen Einschätzungen von Schülertexten werden zwei unterschiedliche Kodierstrategien miteinander vergleichen und hinsichtlich ihrer Eignung in Large-Scale-Assessments untersucht.

Der dritte empirische Beitrag mit dem Titel „Diagnostik der Rechtschreibkompetenz in der Grundschule – Konstruktprüfung mittels Fehler- und Dimensionsanalysen“ (Böhme & Bremerich-Vos, 2009) befasst sich mit der Evaluierung der Bildungsstandards im Bereich der Rechtschreibkompetenz und wählt hierfür einen diagnostischen Ansatz, der sowohl quantitative Aussagen über das globale Kompetenzniveau als auch Aussagen auf der Ebene von qualitativen Fehleranalysen innerhalb der orthografischen Stufe des Rechtschreiberwerbs gestattet.

Der vierte empirische Beitrag trägt den Titel „Methodische Aspekte der Erfassung der Lesekompetenz“ (Böhme & Robitzsch, 2009a) und beleuchtet messmethodische Herausforderungen bei der testdiagnostischen Erfassung und Modellierung der Lesekompetenz. Diese beziehen sich zum einen auf methodische Ansätze zur Bestimmung der Konstruktdimensionalität und zum anderen auf Aspekte des differenziellen Itemfunktionierens in leistungsheterogenen Gruppen.

Der Rahmentext der vorliegenden Arbeit schließt mit einer Zusammenschau der empirischen Beiträge und der kritischen Diskussion aktueller Herausforderungen im Kontext von Bildungsstandards und Large-Scale-Assessments.

Abstract

This doctoral dissertation is based in the field of educational science and explores the assessment of language abilities in German. In my thesis, I address both measurement aspects and didactical considerations in language assessment and present the empirical results of analyses of the National Education Standards for German language at the elementary school level. The thesis consists of four empirical articles and is embedded in a theoretical framework.

The introductory section of the dissertation details the theoretical context and relevant scientific disciplines and defines the central concept of educational competence. I further discuss methodological aspects and address the language competencies described by the National Educational Standards.

The first article, entitled "On the distinction of listening and reading comprehension with the help of characteristics that affect item difficulty" (Böhme, Robitzsch & Busè, 2010), focuses on the Large-Scale-Assessment of listening comprehension and presents results of the analyses of the effect of item characteristics on item difficulty in the development of standards-based assessment tasks.

The second article, entitled "Issues of the evaluation of writing" (Böhme, Bremerich-Vos & Robitzsch, 2009), explores aspects of writing assessment and different evaluation procedures for writing products. An investigation of inter-rater-reliability, as well as a comparison between analytic and holistic coding is conducted and the practicality of both strategies in Large-Scale-Assessment is discussed.

The third article, entitled "The assessment of orthographic competence at elementary level – an evaluation of dimensionality" (Böhme & Bremerich-Vos, 2009), focuses on the study of orthographic competency as defined by the National Educational Standards. The analyses combine quantitative analyses of the general level of competence, as well as qualitative error-analyses at the orthographic level.

The fourth article, entitled "Methodological aspects of reading assessment" (Böhme & Robitzsch, 2009a), illustrates methodological challenges in testing and modeling reading comprehension. The content of this article is twofold: On the one hand, it examines methodological means to identify the dimensionality of constructs and on the other hand, it addresses differential item functioning in groups with dissimilar levels of competency.

The final section of the theoretical framework summarizes the main ideas and findings of the four articles and provides a critical discussion of current challenges in standard-based testing in the context of National Educational Standards and Large-Scale-Assessments.

Inhaltsverzeichnis

Zusammenfassung III
Abstract IV

1 Struktur der vorliegenden Arbeit 1
2 Begriffsklärungen und Vorüberlegungen 4

2.1 Verortung der vorliegenden Arbeit in den relevanten Wissenschaftsdisziplinen 4
2.1.1 Die Deutschdidaktik: Theoretische Grundlage für die Diagnostik sprachlicher Kompetenzen in der empirischen Bildungsforschung 5
Aktuelle deutschdidaktische Arbeitsfelder im Kontext von Bildungsstandards 5
Die Rolle empirischer Forschung in der Deutschdidaktik 8
Perspektiven der Deutschdidaktik 10
2.1.2 Die pädagogisch-psychologische Diagnostik als Bestandteil der empirischen Bildungsforschung 11
2.1.3 Die empirische Bildungsforschung und ihr Verhältnis zur Erziehungswissenschaft 15
2.1.4 Einbettung der vorliegenden Arbeit in die vorgestellten Disziplinen 20
2.2 Die länderübergreifenden Bildungsstandards in Deutschland 21
2.2.1 Hintergründe der Einführung länderübergreifender Bildungsstandards 21
2.2.2 Charakterisierung der länderübergreifenden Bildungsstandards 25
2.3 Der Kompetenzbegriff 28
2.3.1 Sprachwissenschaftliches Begriffsverständnis 29
2.3.2 Psychologisches Begriffsverständnis 32
2.3.3 Kritische Diskussion des kognitionsorientierten Kompetenzbegriffs 34
2.3.4 Der Begriff der Sprachkompetenz 37
2.4 Die Operationalisierung von latenten Kompetenzkonstrukten und messmethodische Grundlagen 39
2.4.1 Überlegungen zur Konstruktvalidierung 39
2.4.2 Die Spezifikation von Messmodellen 42
2.4.3 Konsequenzen der Modellspezifikation am Beispiel der Lesekompetenz 46
2.5 Sprache als Gegenstand wissenschaftlicher Forschung 49
2.6 Exkurs: Sprachentwicklung 51

3 Einführung der in den Bildungsstandards thematisierten sprachlichen Kompetenzen 54

3.1 Die Struktur sprachlicher Kompetenzen 54

3.2 Kompetenzbereich I – Sprechen und Zuhören 62

3.2.1 Aspekte mündlicher Sprachkompetenz 63

3.2.2 Exkurs: Szenisch Spielen 67

3.2.3 Teilbereich Sprechen 68

3.2.4 Teilbereich Zuhören 72

I. Beitrag 1: Zur Abgrenzung des Hörverstehens gegenüber dem Leseverstehen mit Hilfe schwierigkeitsbestimmender Merkmale bei der Entwicklung von Testaufgaben 75

I.1 Theoretische Grundlagen der Konstruktdefinition des Hörverstehens 76

I.1.1 Aspekte der Konstruktoperationalisierung 76

I.1.2 Muttersprachliches versus fremdsprachliches Hörverstehen 78

I.1.3 Konstruktkonfundierung 79

I.1.4 Abgrenzung des Hörverstehens gegenüber dem Leseverstehen 79

I.2. Schwierigkeitsbestimmende Merkmale in der Entwicklung von Testaufgaben 81

I.3 Methode 85

I.3.1 Beschreibung der Personen- und Itemstichprobe 85

I.3.2 Grundlagen der Skalierung 85

I.4 Regressions- und Kommunalitätenanalysen 86

I.5 Empirische Befunde 89

I.5.1 Ergebnisse zum Hörverstehen 89

I.5.2 Vergleichende Ergebnisse zum Leseverstehen 91

I.6 Diskussion und Ausblick 92

I.7 Literatur 95

3.3 Kompetenzbereich II – Schreiben 98

3.3.1 Texte verfassen 98

II. Beitrag 2: Aspekte der Kodierung von Schreibaufgaben 101

II.1 Überblick 103

II.2 Schreibkompetenz – Konstrukt und Erwerb 104

II.3 Schreibkompetenz in den Bildungsstandards und Möglichkeiten ihrer Testung 106

II.3.1 Messung von Schreibkompetenzen 107

II.4 Das Problem der reliablen Beurteilung von Schüleraufsätzen 108
II.4.1 Maße der Interraterreliabilität und Beurteilerübereinstimmung 111
II.5 Varianten der Kodierung 114
II.6 Fragestellungen 116
II.7 Anlage der Untersuchung 116
II.7.1 Testentwicklung und Stichprobenbeschreibung 116
II.7.2 Beispielaufgabe 117
II.7.3 Raterdesign 119
II.7.4 Methodisches Vorgehen 120
II.8 Umsetzung der alternativen Kodierungsvarianten 123
II.8.1 Analytische Kodierung 123
II.8.2 Holistische Kodierung 126
II.9 Ergebnisse 129
II.9.1 Interraterreliabilität und Beurteilerübereinstimmung 129
II.9.2 Vergleich von holistischer und analytischer Kodierstrategie 131
Befunde zum inhaltlichen Bereich 131
Befunde zum allgemein sprachlichen Bereich 134
II.10 Diskussion und Ausblick 138
II.10.1 Methodische Aspekte 139
II.10.2 Implikationen für zukünftige Assessments 140
II.11 Literatur 143

3.3.2 Richtig schreiben 148

III. Beitrag 3: Diagnostik der Rechtschreibkompetenz in der Grundschule – Konstruktprüfung mittels Fehler- und Dimensionsanalysen 150

III.1 Überblick 151
III.2 Rechtschreibkompetenz 152
III.3 Erwerb der Rechtschreibkompetenz 153
III.4 Rechtschreibkompetenz in den Bildungsstandards und Möglichkeiten ihrer Testung 154
III.5 Diagnostische Möglichkeiten im Bereich der Rechtschreibung 156
III.6 Kontextualisierung und Präzisierung der Fragestellungen 158
III.7 Anlage der Untersuchung 159
III.7.1 Testentwicklung 159
Diagnostische Kategorien 160
Didaktische Rationale der Auswahl von Testwörtern 161
Beschreibung der Testinstrumente 162

III.7.2 Datengrundlage 163
III.7.3 Statistische Analysen 164
III.8 Ergebnisse 165
III.8.1 Deskriptive Befunde 165
III.8.2 Geschlechts- und Jahrgangsdifferenzen 167
III.8.3 Psychometrische Kennwerte und Dimensionsanalysen 170
III.9 Diskussion 175
III.10 Literatur 179

3.4 Kompetenzbereich III – Lesen 184
3.4.1 Der Textbegriff 184
3.4.2 Lesekompetenz und kognitionspsychologische Grundlagen des Leseverstehens 186
3.4.3 Die Bildungsstandards im Kompetenzbereich Lesen und Möglichkeiten ihrer Testung 189

IV. Beitrag 4: Methodische Aspekte der Erfassung der Lesekompetenz ... 191

IV.1 Einführung und Überblick 192
IV.2 Aspekte der Dimensionalität der Lesekompetenz im Primarbereich 193
IV.2.1 Die Konstruktdimensionalität und -stabilität der Lesekompetenz 193
Methodische Vorüberlegungen zur Dimensionalitätsprüfung 194
Empirische Befunde zur Konstruktdimensionalität der Lesekompetenz im Deutschen 199
Stabilität der Konstruktdimensionalität über den Erwerbsprozess 201
IV.2.2 Fragestellung 202
IV.2.3 Methode 202
Beschreibung der Personen- und Itemstichprobe 202
Methodische Zugänge zur Analyse der Konstruktdimensionalität 202
Nichtparametrische Ansätze 204
Parametrische Ansätze 205
IV.2.4 Befunde der Dimensionalitätsanalysen 207
Befunde nichtparametrischer Ansätze (DETECT) 208
Untersuchung der Dimensionalität auf Aufgabenebene mittels parametrischer Limited Information-Ansätze 209
Befunde parametrischer Ansätze unter Rückgriff auf Full-Information-Maximum-Likelihood-Verfahren 210
IV.2.5 Zusammenfassung und Diskussion 212
IV.3 Differenzielles Itemfunktionieren in der dritten und vierten Jahrgangsstufe 213
IV.3.1 Bildungspolitische Ausgangssituation 213
IV.3.2 Methode 215

Charakterisierung der Personenstichproben ... 215
Methodische Grundlagen der Verlinkung der Jahrgangsstufen drei und vier ... 216
Differential Item Functioning und Differential Testlet Functioning (DIF, DTF) ... 217

IV.3.3 Befunde zur Verlinkung der Jahrgangsstufen drei und vier sowie zum differenziellen Itemfunktionieren ... 220
Befunde zur Verlinkung beider Jahrgangsstufen ... 220
Befunde zu verschiedenen Methoden der vertikalen Skalierung ... 221
Befunde zum differenziellen Itemfunktionieren zwischen den Klassenstufen ... 223

IV.3.4 Zusammenfassung und Diskussion ... 226
IV.4 Zusammenführung der Befunde und Ableitung von Empfehlungen ... 227
IV.5 Literatur ... 231

3.5 Kompetenzbereich IV – Sprache und Sprachgebrauch untersuchen ... 238
3.5.1 Das Kompetenzkonstrukt „Sprache und Sprachgebrauch untersuchen“ aus Sicht der Deutschdidaktik ... 239
3.5.2 Überlegungen zur Konstruktoperationalisierung ... 242
3.5.3 Fazit ... 247

4 Abschließende Diskussion ... 248

4.1 Zusammenschau der empirischen Beiträge ... 248
4.1.1 Zur Untersuchung des Hörverstehens mittels schwierigkeitsbestimmender Merkmale ... 248
4.1.2 Zur Kodierung von Schülertexten im Kompetenzbereich Schreiben ... 250
4.1.3 Zur Diagnostik der Rechtschreibkompetenz ... 254
4.1.4 Zu methodischen Aspekten der Erfassung der Lesekompetenz ... 256
4.2 Einschätzung der Bildungsstandards für das Fach Deutsch im Primarbereich ... 259
4.2.1 Die Umsetzung der Bildungsstandards für das Fach Deutsch in Testaufgaben ... 259
4.2.2 Konsequenzen der Bildungsstandards und regelmäßiger Testdurchführungen für den Deutschunterricht ... 263
4.2.3 Fachspezifische vs. fächerübergreifende Kompetenzen ... 265
4.3 Nutzen und Nutzbarmachung von Kompetenzdiagnostik ... 266
4.3.1 Kompetenzdiagnostik als Werkzeug der Qualitätsentwicklung im Bildungswesen ... 266
4.3.2 Erwartungsdivergenz: System- vs. Individualdiagnostik ... 268
4.3.3 Rückmeldeverfahren: Die (unzureichende) Nutzbarmachung der Ergebnisse von Kompetenzdiagnostik ... 270
4.4 Teaching to the Test ... 275
4.5 Zukunftsperspektiven der Erziehungswissenschaft ... 279
4.6 Kompetenzentwicklung ... 287

4.6.1 Das Problem der Modellierung von Kompetenzentwicklungstrends 287
4.6.2 Methodische Defizite im Rahmen der Vergleichsarbeiten im Fach Deutsch in den Jahrgangsstufen 3 und 8 291
4.6.3 Die Entwicklung und Etablierung von Kompetenzmodellen und Kompetenzentwicklungsmodellen 293

5 Literatur 296

1 Struktur der vorliegenden Arbeit

Die vorliegende Arbeit widmet sich der empirischen Analyse sprachlicher Kompetenzen im *Primarbereich* und tut dies vor dem Hintergrund der länderübergreifenden *Bildungsstandards* für das Fach *Deutsch* (vgl. KMK, 2004, 2005a, 2005b). Länderübergreifende Bildungsstandards für das Fach Deutsch liegen seit Beginn des neuen Jahrtausends für die Primarstufe sowie für die Sekundarstufe I, und hier für den Hauptschulabschluss sowie den Mittleren Schulabschluss (MSA) vor. Seit den Jahren 2003 bzw. 2004 haben die Bildungsstandards verbindlichen Charakter. Gegenwärtig werden ferner Bildungsstandards für das Fach Deutsch in der Sekundarstufe II erarbeitet. Diese Dissertation befasst sich in ihren theoretischen Betrachtungen vorwiegend und in den empirischen Analysen ausschließlich mit den Bildungsstandards für das Fach Deutsch im Primarbereich (vgl. KMK, 2005a).

Priorität hat hierbei die Frage, inwieweit es gelungen ist, die Bildungsstandards als Grundlage einer Kompetenzmessung so zu operationalisieren, dass sie einerseits den theoretisch und praktisch relevanten, deutschdidaktischen sowie erziehungswissenschaftlichen Fundamenten und andererseits den Maßstäben der pädagogisch-psychologischen Diagnostik im Rahmen von *Large-Scale-Assessments* genügen.

Diese Dissertation besteht aus vier Einzelbeiträgen, die sich unterschiedlichen empirischen Fragestellungen zur Diagnostik sprachlicher Kompetenzen widmen sowie einer theoretisch orientierten Rahmung, welche die Einzelbeiträge integriert und umschließt.

Hierfür erfolgt in Abschnitt 2.1 des Rahmentextes zunächst eine Einordnung der vorliegenden Arbeit in die wissenschaftlichen Referenzdisziplinen. Zu verorten sind die in den empirischen Beiträgen dargestellten Studien und Überlegungen im Arbeitsfeld der empirischen Bildungsforschung. Wesentliche Impulse stammen aus der Deutschdidaktik (Abschnitt 2.1.1), der pädagogisch-psychologischen Diagnostik als Bestandteil der Pädagogischen Psychologie (Abschnitt 2.1.2) sowie der Erziehungswissenschaft (Abschnitt 2.1.3).

Im Anschluss an die wissenschaftliche Verortung dieser Dissertation werden im Abschnitt 2.2 die Bildungsstandards für das Fach Deutsch im Primarbereich eingeführt, in ihren historischen Entstehungszusammenhang eingebettet und hinsichtlich ihrer spezifischen Merkmale charakterisiert.

In Abschnitt 2.3 folgt eine ausführliche Diskussion des für diese Arbeit zentralen Kompetenzbegriffs unter Bezugnahme auf dessen sprachwissenschaftliche und psychologische Wurzeln. Anschließend wird in Abschnitt 2.4 auf die Hintergründe und Konsequenzen der Operationalisierung von latenten Kompetenzkonstrukten sowie methodische Grundlagen und Vorüberlegungen eingegangen.

In den nachfolgenden Abschnitten 2.5 und 2.6 wird Sprache als zentraler Gegenstand dieser Arbeit in den Blick genommen und als Inhalt wissenschaftlicher Forschung thematisiert. Die frühkindliche Sprachentwicklung wird in einem Exkurs dargestellt und als Erklärungsgrundlage späterer empirischer Befunde eingeführt.

Das zentrale Kapitel 3 beginnt mit einer kurzen Betrachtung der in den Bildungsstandards für das Fach Deutsch ausgewiesenen sprachlichen Kompetenzen und beleuchtet in Abschnitt 3.1 ihre strukturellen Zusammenhänge. Nachfolgend werden die einzelnen Kompetenzbereiche eingehender untersucht, wobei die in den Bildungsstandards für das Fach Deutsch (KMK, 2004, 2005a, 2005b) gewählte Reihenfolge beibehalten wird.

Somit wird in Abschnitt 3.2 zunächst der Kompetenzbereich *Sprechen und Zuhören* näher betrachtet, wobei in Vorbereitung des ersten Aufsatzes der Schwerpunkt der theoretischen Auseinandersetzung auf dem Teilbereich *Zuhören* liegt. Es folgt der empirische Beitrag „Zur Abgrenzung des Hörverstehens gegenüber dem Leseverstehen mit Hilfe schwierigkeitsbestimmender Aufgabenmerkmale bei der Entwicklung von Testaufgaben“. Dieser Aufsatz beschäftigt sich mit der Operationalisierung des Zuhörens für Testungen im Rahmen von Large-Scale-Assessments und verbindet hierfür theoretische Überlegungen zur Konstruktoperationalisierung mit der Definition von schwierigkeitsbestimmenden Merkmalen. Die empirischen Analysen beziehen sich auf Studien zur Pilotierung und Normierung der Bildungsstandards im Fach Deutsch für den Primarbereich.

Abschnitt 3.3 widmet sich dem Kompetenzbereich *Schreiben* und untergliedert diesen in die Teilbereiche *freies Schreiben (Texte verfassen)* und *Rechtschreibung (richtig schreiben)*. Zu jedem dieser beiden Teilbereiche liegt ein eigener empirischer Beitrag vor. Der Aufsatz „Aspekte der Kodierung von Schreibaufgaben“ verfolgt das Ziel, fachdidaktische Überlegungen zum Konstrukt der Schreibkompetenz mit diagnostischen Fragen der Operationalisierung und der Bewertung von Schreibprodukten zu verbinden. Aus fachdidaktischer sowie diagnostischer Sicht ist von besonderem Interesse, welche Komponenten von Schreibkompetenz unterschieden werden können, da dies Implikationen für verschiedene Möglichkeiten der Bewertung birgt. In diesem Zusammenhang sind ferner Aspekte der Interraterreliabilität, also der Übereinstimmung zwischen verschiedenen Bewertern bei der Beurteilung von Schreibprodukten relevant. Der sich anschließende empirische Beitrag mit dem Titel „Diagnostik der Rechtschreibkompetenz in der Grundschule – Konstruktprüfung mittels Fehler- und Dimensionsanalysen“ integriert quantitative Aussagen über das globale Kompetenzniveau von Grundschulkindern in diesem Bereich mit qualitativen Fehleranalysen innerhalb der orthografischen Stufe des Rechtschreiberwerbs.

In Abschnitt 3.4 folgt eine vertiefte Auseinandersetzung mit dem Kompetenzbereich *Lesen – mit Texten und Medien umgehen*. Hierbei wird zunächst das Konstrukt Lesekompetenz charakterisiert und der Textbegriff untersucht. Der anschließende empirische Beitrag „Methodische Aspekte der Erfassung der Lesekompetenz" befasst sich mit methodischen Herausforderungen und Besonderheiten wie parametrischen und nichtparametrischen Verfahren zur Bestimmung der Konstruktdimensionalität und dem differenziellen Funktionieren von Items in leistungsheterogenen Gruppen.

Der Kompetenzbereich *Sprache und Sprachgebrauch untersuchen* erfährt in Abschnitt 3.5 eine ausführliche Würdigung, wobei wiederum sowohl deutschdidaktische als auch diagnostische Aspekte thematisiert werden.

Den Abschluss der theoretischen Rahmung bildet Kapitel 4, welches zum einen die kritische Diskussion der in diese Dissertation einbezogenen empirischen Beiträge umfasst und zum anderen weiterführende Debatten um aktuelle Probleme und Herausforderungen in der empirischen Bildungsforschung aufgreift. Hierbei handelt es sich um eine Einschätzung der Bildungsstandards für das Fach Deutsch im Primarbereich aus unterschiedlichen Perspektiven, die unzureichende Nutzbarmachung der Ergebnisse von Kompetenzdiagnostik in Rückmeldeverfahren sowie die Kompetenzdiagnostik als Werkzeug der Qualitätsentwicklung im Bildungswesen. Ferner wird das Problem des *Teaching to the Test* diskutiert und das kritische Verhältnis der Erziehungswissenschaft zur quantitativ-empirisch arbeitenden Bildungsforschung beleuchtet. Die Diskussion schließt mit einem Ausblick auf die Trendmodellierung von Kompetenzentwicklung, welche eine zentrale Herausforderung für die empirische Bildungsforschung der nächsten Jahre und Jahrzehnte darstellen wird.

2 Begriffsklärungen und Vorüberlegungen

2.1 Verortung der vorliegenden Arbeit in den relevanten Wissenschaftsdisziplinen

Die Diagnostik sprachlicher Kompetenzen stellt eine komplexe Herausforderung dar, für deren Bewältigung auf Erkenntnisse aus verschiedenen Wissenschaftsdisziplinen zurückgegriffen werden kann und muss.

Hierbei sind für die vorliegende Arbeit in einem ersten Schritt Aussagen zu den theoretischen Grundlagen, präzise Definitionen der jeweiligen sprachlichen Kompetenzkonstrukte und ihre Verortung in Kompetenzmodellen von Belang, welche primär von der *Deutschdidaktik* sowie in geringerem Umfang von der Sprachpsychologie beziehungsweise der Psycholinguistik beigesteuert werden.

Erkenntnisse der *pädagogisch-psychologischen Diagnostik* liefern die entscheidenden Voraussetzungen für die Messung eines Konstrukts und gestatten so eine passgenaue und zielführende Operationalisierung der in dieser Dissertation betrachteten sprachlichen Kompetenzkonstrukte.

Die heute gesellschaftlich relevanten, bildungsbezogenen Fragestellungen der vorliegenden Arbeit knüpfen an das traditionelle Arbeitsfeld der *Erziehungswissenschaft* an, welche somit eine weitere Referenzdisziplin darstellt.

Die *empirische Bildungsforschung* selbst bildet den primären Bezugsrahmen dieser Dissertation und stellt zum einen spezifisches methodisches Wissen für den Umgang mit Large-Scale-Daten zur Verfügung, zum anderen bestimmt sie den Kontext, innerhalb dessen die hier betrachteten Fragestellungen relevant und die gewonnenen Erkenntnisse zu interpretieren sind.

In den folgenden Abschnitten sollen die für diese Dissertation wesentlichen Referenzdisziplinen gewürdigt werden. Hierbei wird zunächst jeweils eine allgemeine Charakterisierung vorgenommen, bevor in Abschnitt 2.1.4 konkrete Bezüge zu den in dieser Dissertation untersuchten Forschungsfragen hergestellt werden.

2.1.1 Die Deutschdidaktik: Theoretische Grundlage für die Diagnostik sprachlicher Kompetenzen in der empirischen Bildungsforschung

Aktuelle deutschdidaktische Arbeitsfelder im Kontext von Bildungsstandards

Zentraler Gegenstand der Deutschdidaktik ist – beziehungsweise war lange Zeit – die Entwicklung neuer Konzeptionen für den Deutschunterricht. Hierbei kommt oftmals die starke fachwissenschaftliche Verankerung der Disziplin zum Tragen, so dass in aller Regel Gegenstandstheorien der Linguistik und Literaturwissenschaft so bearbeitet werden, dass sie im schulischen Alltag didaktisch nutzbar sind (vgl. bspw. Bremerich-Vos, 2002; Kammler & Knapp, 2002). Vertreter der Deutschdidaktik verstehen ihre Disziplin also häufig als eine Wissenschaft vom *fachbezogenen* Lehren und Lernen. Die Fokussierung auf fachwissenschaftliche, also linguistische und literaturwissenschaftliche Grundlagen birgt mitunter die Gefahr einer Einengung des eigenen wissenschaftlichen Horizonts, da verschiedene primär *pädagogische* Fragen auf diese Weise nicht in den Blick geraten können, obwohl diese für das Geschehen im (Deutsch-)Unterricht von wesentlicher Bedeutung sind.

Die öffentliche und politische Rezeption der Ergebnisse der PISA-Studie des Jahres 2000 (*Programme for International Student Assessment*; Baumert et al., 2001, 2002) haben auch innerhalb der Deutschdidaktik zu einer Reflexion bezüglich des Selbstverständnisses der Disziplin und zu einer partiellen Neuorientierung geführt (vgl. Abraham, Bremerich-Vos, Frederking & Wieler, 2003). Hinzukommt, dass durch die Einführung der länderübergreifenden Bildungsstandards für das Fach Deutsch (KMK, 2004, 2005a, 2005b) neue Erwartungen an die Deutschdidaktik herangetragen werden und sich diese veränderten Arbeitsschwerpunkten gegenübersieht.

Die Wahrnehmung, inwieweit Vertreterinnen und Vertreter der Deutschdidaktik in die Entwicklung der Bildungsstandards, der standardbasierten Test- und Unterrichtsaufgaben sowie der hierfür ausschlaggebenden Kompetenzmodelle eingebunden waren beziehungsweise werden, divergiert innerhalb der Disziplin beträchtlich. Während beispielsweise Speck-Hamdan (2007) davon spricht, dass „das IQB[1] zunächst den Kontakt zwischen Diagnostik und Fachdidaktik sowie zur Praxis her[stellte] und […] beide Gruppen in die Konstruktion von Aufgaben ein[band]“ (S. 91f.), vertreten andere die Ansicht, dass „[…] Bildungspolitik und Schuladministration […] ohne nennenswerte Beteiligung der Deutschdidaktik Bildungsstandards und Kompetenzen formulieren, implementieren und evaluieren […]“ (Steinbrenner, 2007, S. 5). Die Divergenz der Perspektiven scheint insbesondere dadurch begründet, dass in die

[1] Institut zur Qualitätsentwicklung im Bildungswesen

Überführung der länderübergreifenden Bildungsstandards in Test- und Unterrichtsaufgaben sowie in die Entwicklung von Kompetenz- und Kompetenzstufenmodellen nur eine Auswahl deutschdidaktischer Experten partizipativ eingebunden werden konnte. Die Beteiligung der Deutschdidaktik ist somit notwendigerweise exemplarisch und insofern selektiv, als eine interdisziplinäre Kooperation die grundsätzliche Bereitschaft für eine solche Zusammenarbeit erfordert. Dass die Disziplin als solche von den aktuellen bildungspolitisch initiierten Bemühungen der Qualitätsentwicklung ausgegrenzt wurde, trifft aus meiner Sicht aber nicht zu.

Wichtige Funktion jeder Fachdidaktik ist die normative Setzung dessen, was Gegenstand des Unterrichtsfaches ist, was also von den Lehrkräften unterrichtet und auf Seiten der Schülerinnen und Schüler verinnerlicht werden soll. Die Bildungsstandards für das Fach Deutsch im Primarbereich (KMK, 2005a) sowie für den Hauptschulabschluss und den Mittleren Schulabschluss (KMK, 2004, 2005b) übernehmen diese Funktion der normativen Regulierung nur zu einem gewissen Teil. Dies erklärt sich dadurch, dass sie output-orientiert definieren, was Kinder und Jugendliche zu einem bestimmten Zeitpunkt in ihrer Schullaufbahn gelernt haben sollen, ohne dabei festzulegen, auf welchem Weg dieses Ziel zu erreichen ist (vgl. Abschnitt 2.2.2; Köller, 2008, 2010). Den Lehrkräften vor Ort bleibt also freigestellt, mit welchen Unterrichtskonzepten sie arbeiten, welche Lehrbücher und Unterrichtsmaterialien sie einsetzen, welche literarischen Werke sie konkret behandeln etc. Auch die Fragen, wie kompetenzorientierte Unterrichtskonzepte charakterisiert werden können und wie ein kompetenzorientierter Unterricht in den verschiedenen Domänen konkret gestaltet werden kann und sollte, bedürfen der Klärung (vgl. Abraham, Baurmann, Feilke, Kammler & Müller, 2007). Diese Leerstelle zu schließen, ist gegenwärtig eine wichtige Aufgabe der Fachdidaktiken und somit auch der Deutschdidaktik. Einen ersten Schritt auf diesem Weg stellt eine von Deutschdidaktikern in Kooperation mit dem IQB erarbeitete Sammlung und Erläuterung von bildungsstandardbasierten Lernaufgaben, Unterrichtsanregungen und Fortbildungsideen dar, die explizit dem Ziel eines kompetenzorientierten und bildungsstandardbasierten Unterrichts verpflichtet ist (vgl. Bremerich-Vos, Granzer, Behrens & Köller, 2009).

In der Auseinandersetzung mit den länderübergreifenden Bildungsstandards für das Fach Deutsch rücken außerdem zwei weitere Arbeitsfelder in den Mittelpunkt der deutschdidaktischen Auseinandersetzung: zum einen die Erarbeitung von *Kompetenzmodellen* und zum anderen die *Aufgabendidaktik* (vgl. Böhnisch, 2008; Frederking et al., 2003; Köster & Lindauer, 2008).

„*Kompetenzmodelle* konkretisieren Inhalte und Stufen der allgemeinen Bildung. Sie formulieren damit eine *pragmatische Antwort auf die Konstruktions- und Legitimationsprobleme traditioneller Bildungs- und Lehrplandebatten*“ (Hervorhebungen im Original; Klieme et al., 2007, S. 9). Sie bieten den Lehrkräften ein Referenzsystem für ihr professionelles Handeln. Außerdem fließt in

Kompetenzmodellen fachdidaktisches und pädagogisch-psychologisches Wissen zusammen (Klieme et al., 2007). Böhnisch (2008, S. 5) bezeichnet die Entwicklung von Kompetenzmodellen als eine der Kernaufgaben der Fachdidaktik. Auch Frederking und Kollegen (2003) postulieren: „Die Entwicklung von wissenschaftlich begründeten und haltbaren, domänenspezifischen Kompetenzmodellen sowie die Anwendung gestufter Verfahren zur Definition von Kompetenzskalen und zur Kalibrierung standardisierter Aufgaben ist das genuine Aufgabenfeld der Fachdidaktiken, hier: der Sprachdidaktiken“ (S. 2). Als zentrales Problem einer empirisch basierten Entwicklung von Kompetenzmodellen wird diskutiert, dass verschiedene relevante Erfahrungen und Fähigkeiten nicht ohne weiteres mit dem Kompetenzbegriff erfasst werden können und somit in den gegenwärtig eingesetzten Testinstrumenten und Kompetenzmodellen unterrepräsentiert sind. Spinner (2008) mahnt in diesem Kontext an, „das Bewusstsein für jene Bildungsaspekte, die sich der Formulierung als kompetenzorientierte Standards nicht fügen wollen, offen zu halten“ (S. 322). Auch wird innerhalb der Deutschdidaktik über den Auflösungsgrad von Kompetenzmodellen diskutiert. Da die gegenwärtig in der empirischen Bildungsforschung verwendeten Kompetenzmodelle als zu grobmaschig eingeschätzt werden, besteht auf Seiten der Deutschdidaktik verschiedentlich der Wunsch nach einer feineren, „domänenbezogenen Aufschlüsselung“ (Abraham et al., 2007, S. 7). Dass eine Kompetenzmodellierung für Zwecke des Bildungsmonitorings aber möglicherweise auf einen anderen – und zwar gröberen – Auflösungsgrad abzielen sollte als die Modellierung sprachlicher Kompetenzen für die unterrichtliche Praxis, die dem Kompetenzaufbau dient, liegt nahe.

Im Zuge der Überführung der Bildungsstandards in *Aufgaben* – aber auch als klassisches Werkzeug der unterrichtlichen Praxis – hat sich die wissenschaftliche Aufgabendidaktik, also die Erarbeitung, Analyse und Reflexion von Lern- und Leistungsaufgaben, als weiterer wesentlicher Gegenstand der gegenwärtigen deutschdidaktischen Arbeit und Forschung etabliert (vgl. bspw. Köster, 2008, 2010; Köster & Lindauer, 2008). Aktuelle Arbeiten beziehen sich beispielsweise auf die Unterscheidung von Lern- und Leistungsaufgaben (vgl. bspw. Abraham & Müller, 2009; Köster, 2008, 2010; Speck-Hamdan, 2007), die Diskussion von schwierigkeitsbestimmenden Merkmalen von Texten sowie Aufgaben (vgl. bspw. Böhme, Robitzsch & Busè, 2010) oder die Standardkompatibilität und Kompetenzorientierung von Aufgaben in Lehrwerken (vgl. bspw. Leubner, 2005; Winkler, 2005). In der Diskussion um Leistungs- und Lernaufgaben äußert Speck-Hamdan (2007, S. 91), dass erst durch die Entwicklung von Illustrationsaufgaben fachdidaktisches Denken in den Prozess der Qualitätssteigerung durch Bildungsstandards einfließen könne. Auch in den relevanten Fachzeitschriften für Grundschulpädagogik sowie den Deutschunterricht werden Fragen der Aufgabenkonstruktion und des kompetenzorientierten

Unterrichtens regelmäßig vertieft behandelt und für den schulischen Einsatz aufbereitet (vgl. bspw. Grundschule 5-2009; Praxis Deutsch 214, 2009).

Die Rolle empirischer Forschung in der Deutschdidaktik

Traditionell wurden in der Deutschdidaktik neu entwickelte Unterrichtskonzepte vor der Implementation in der Praxis nur selten einer empirischen Evaluation hinsichtlich ihrer Wirksamkeit und ihres Nutzens gegenüber der bisherigen Unterrichtspraxis unterzogen (vgl. Kammler & Knapp, 2002). Somit fehlte weitgehend die Rückwirkung empirischer Befunde auf die Theorieentwicklung, etwa im Sinne einer Falsifikation aufgestellter Hypothesen, wie dies in stärker empirisch ausgerichteten Forschungsdisziplinen üblich ist. Aus diesem Grund wurden Mängel in den Unterrichtskonzeptionen oftmals nicht vorab in stichprobenbasierten, eventuell sogar experimentellen Studien offenbar, sondern erst bei ihrer Einführung. Dies hatte zur Folge, dass sich etwaige Unzulänglichkeiten der Neuerungen unmittelbar auf das Lehren und Lernen einer großen Anzahl von Lehrkräften sowie Schülerinnen und Schülern auswirkte. Kammler und Knapp (2002) bemerken zu dieser Situation: „Wenn man genau hinsieht, gibt es fast keine gängige Praxis im Deutschunterricht, die durch empirische Forschung wirklich abgesichert wäre" (S. 6). Zu Recht wurde der Deutschdidaktik aufgrund dieser Arbeitsweise mitunter eine gewisse Praxisferne vorgeworfen. Bremerich-Vos (2002) äußert in diesem Zusammenhang: „Soweit die Deutschdidaktik eine auf ‚Praxis' zielende Wissenschaft ist, hat sie u. a. das Ihre dafür zu tun, dass sie mit den im Unterricht handelnden Lehrenden in Kontakt bleibt. Es gibt eine ganze Reihe von Anzeichen dafür, dass dieser Kontakt – vorsichtig gesagt – sehr brüchig ist" (S. 26).

Im Verlauf des letzten Jahrzehnts wurde daher in der deutschdidaktischen Wissenschaftsgemeinschaft die Wahl der Forschungsmethoden intensiv diskutiert und ein Mangel an empirischen Arbeiten konstatiert (vgl. bspw. Bremerich-Vos, 2002; Groeben, 2005). Hinsichtlich der Konfrontation mit der Forderung nach einer belegbaren Lernwirksamkeit des Deutschunterrichts bemerkt Groeben (2005): „Die Gesellschaft hat ein Anrecht auf die empirisch systematische Überprüfung didaktischer Wirkungsbehauptungen, das sie auch zunehmend durchzusetzen trachtet" (S. 9).

In diesem Kontext wird allerdings verschiedentlich die Ansicht vertreten, dass deutschdidaktische Forschungsgegenstände aufgrund ihrer hohen Komplexität einer empirischen Erforschung nicht in der erforderlichen Weise zugänglich sind. Aus der Perspektive der Deutschdidaktik sind „[…] diverse Kompetenzen, die der Deutschdidaktik traditionell am Herzen liegen, nur schwer oder überhaupt nicht messbar" und werden aufgrund dieser Tatsache

beispielweise bei der Entwicklung von empirisch basierten Kompetenzmodellen zunehmend in den Hintergrund gedrängt (Böhnisch, 2008, S. 5).

Aus Sicht der empirischen Bildungsforschung ist die Einschätzung, dass die deutschdidaktischen Forschungsgegenstände meist zu komplex wären, um sie empirisch so zu untersuchen, dass objektive und übertragbare Ergebnisse resultieren könnten (Kammler & Knapp, 2002, S. 8), allerdings nur bedingt korrekt und trifft in dieser Verallgemeinerung nicht zu. Vielmehr besteht die Herausforderung darin, Messinstrumente, Forschungsdesigns und Analysemethoden so zu wählen beziehungsweise neu zu entwickeln, dass auch die hoch komplexen Forschungsgegenstände der Deutschdidaktik einer empirischen Untersuchung zugänglich gemacht werden können. Dies scheint in erster Linie eine Frage der zu investierenden Ressourcen, nicht aber der prinzipiellen (Un-)Möglichkeit zu sein. In diesem Zusammenhang sollte auch deutlich werden, dass es sich bei empirischer Forschung keinesfalls um ausschließlich quantitativ-empirische Forschung handeln muss. Insbesondere die sehr aufschlussreiche mathematikdidaktische Unterrichtsforschung im Rahmen der TIMSS-Videostudie (vgl. für einen Überblick Klieme & Rakoczy, 2008) verdeutlicht, dass erst die Kombination und Integration verschiedener Forschungsmethoden und -ansätze eine derartige Vielfalt interessanter und relevanter Resultate hervorbringen konnte.

Verschiedene Vertreter der empirischen Bildungsforschung, wie beispielsweise Klieme und Rakoczy (2008, S. 222f.), plädieren für eine enge Kooperation zwischen empirischer Bildungsforschung und Fachdidaktik bei der Modellierung und Messung von Kompetenzen sowie der empirischen Erforschung von Prozessmerkmalen eines erfolgreichen Unterrichts. Sie betonen, dass fachdidaktische und fachliche Expertise für die Entwicklung von bildungsstandardbezogenen Kompetenzmodellen und Testinstrumenten unerlässlich ist. Klieme und Rakoczy (2008, S. 235) weisen aber auch explizit darauf hin, dass sich eine fachdidaktisch fundierte Unterrichtsforschung nicht auf die Modellierung von Kompetenzen beschränken darf, sondern über diese hinausgehen muss. Auch Lipowsky (2007) weist in seiner Auseinandersetzung mit der nationalen und internationalen Erforschung von Unterrichtsqualität in der Grundschule darauf hin, dass künftige Unterrichtsforschung schulische Lernprozesse „unter einer verstärkt fachdidaktischen Perspektive“ (S. 47) analysieren müsse.

Dass empirische Untersuchungen nützliche Hinweise für die deutschdidaktische Arbeit liefern können, belegen auch literaturdidaktische Beiträge der letzten Jahre, die sich um Aufgabenstellungen und Unterrichtsanalysen zu der Kurzgeschichte „Brudermord im Altwasser“ von Georg Britting ranken (vgl. bspw. Kämper-van den Boogaart & Pieper, 2008; Winkler, 2005).

Winkler (2005, S. 194) konnte zeigen, dass der Lernerfolg der Schülerinnen und Schüler stets und zu ganz wesentlichen Anteilen von der jeweiligen Lehrkraft und ihrer

fachwissenschaftlichen sowie didaktischen Expertise determiniert wird. In diesem Zusammenhang konnte Winkler ferner beobachten, dass ein stark lehrergelenkter Instruktionsunterricht für die Vermittlung eines konkreten, kognitiv-analytischen Textverständnisses ausgesprochen zielführend sein kann und im Vergleich mit einem weniger stark lenkenden Unterrichtsstil sogar eine größere Deutungstiefe sowie eine größere terminologische Präzision und eine bessere Bündelung der Ergebnisse zu erzielen vermag (vgl. Winkler, 2005, S. 195). Dieser Befund führte auf dem 15. Symposion Deutschdidaktik im Jahr 2004 zu einer angeregten Diskussion, da ein stark gesteuerter Instruktionsunterricht innerhalb der Disziplin als unpopulär gilt. Gemessen an den empirischen Befunden scheint die pauschale Ablehnung eines solchen Unterrichtsstils aber nicht gerechtfertigt, möglicherweise ist sogar ein Überdenken bislang vehement vertretener Präferenzen angezeigt. Diese beispielhaft herausgegriffene Untersuchung ist geeignet zu verdeutlichen, dass ein empirischer Zugang zweckdienlich sein kann, um begründete Zweifel an einer weit verbreiteten Annahme aufkeimen zu lassen und auf diese Weise auch eine theoretische Weiterentwicklung innerhalb der Deutschdidaktik anzustoßen.

Perspektiven der Deutschdidaktik

Insgesamt scheint sich die Deutschdidaktik in einer Phase der (partiellen) Neuorientierung zu befinden, die sich auf verschiedene Aspekte bezieht. Zunächst betrifft sie die Öffnung verschiedener Vertreter der Deutschdidaktik für empirisches Arbeiten, wobei neben den vertrauten qualitativen und hermeneutischen Methoden auch zunehmend quantitative Zugänge in den Blick genommen werden.

Spezifisch für die Deutschdidaktik war beziehungsweise ist die tradierte Trennung von Sprach- und Literaturdidaktik. Diese schlägt sich nicht nur personell nieder, sondern birgt auch Implikationen für die jeweils vorherrschenden wissenschaftlichen Fragestellungen und Forschungsmethoden. Mitunter wird erwogen, ob die Etablierung der länderübergreifenden Bildungsstandards für das Fach Deutsch (KMK, 2004, 2005a, 2005b) sowie die Hinwendung zu empirischen Forschungsmethoden eine Chance für eine stärkere Integration dieser beiden Felder bergen könnte (vgl. Kammler & Knapp, 2002; Ossner, 2006b). Eine solche Möglichkeit wird allerdings insbesondere von Seiten der Literaturdidaktik kritisch eingeschätzt, da befürchtet wird, dass der pragmatische und instrumentelle Aspekt von Sprache in seiner Bedeutung zu Lasten der Literatur zunimmt (vgl. Steinbrenner, 2007, S. 7).

Ferner zeigen sich viele Vertreter der Deutschdidaktik offen für die Kooperation mit Wissenschaftsdisziplinen, die in der Vergangenheit eher als rivalisierend charakterisiert wurden.

Bremerich-Vos (2002) äußert in diesem Zusammenhang: „Wir DeutschdidaktikerInnen tun m. E. gut daran, […] eingedenk des Umstands, dass eine ganze Reihe von Kernthemen unserer Disziplin längst andernorts empirisch bearbeitet wird, den Blick über den deutschdidaktischen Zaun zu riskieren“ (S. 27). Ebenso wie in der Erziehungswissenschaft werden aber auch innerhalb der Deutschdidaktik gelegentlich Stimmen laut, die der Vorstellung einer „feindlichen Übernahme“ verhaftet scheinen und einer interdisziplinären Kooperation kritisch gegenüberstehen, da ein Klima der Fremdbestimmung herrsche (Steinbrenner, 2007, S. 5). Dass, wie von Steinbrenner (2007, S. 5) postuliert, in der gegenwärtigen empirischen Erforschung sprachlicher Kompetenzen allerdings tatsächlich eine Fokussierung auf *fremde* Konstrukte vorherrschen soll, verwundert. In der Tat findet in der aktuellen bildungsstandardbasierten Forschung und Testentwicklung eine Beschränkung in zweifacher Hinsicht statt. Diese betrifft zum einen das Spektrum der untersuchten sprachlichen Kompetenzen (bspw. Aussparung des Sprechens) sowie die Breite der beschriebenen Kompetenzkonstrukte (bspw. Fokussierung auf kognitive Aspekte) und zum anderen die Differenziertheit der Diagnose (bspw. Produkt- statt Prozessdiagnostik, kaum Einsatz von *Partial-Credit-Modellen*). Dennoch kommt die Frage auf, welche Konstrukte der Deutschdidaktik *eigen* sind, wenn nicht beispielsweise der schulische Erwerb des Lesens und Schreibens, wie er aktuell im Rahmen der Bildungsstandards für das Fach Deutsch untersucht wird.

2.1.2 Die pädagogisch-psychologische Diagnostik als Bestandteil der empirischen Bildungsforschung

Pädagogisch-psychologische Diagnostik dient der Entscheidungsfindung bei bildungsbezogenen Fragen und stellt somit ein wichtiges Element der Pädagogischen Psychologie dar. Für die vorliegende Arbeit bildet die pädagogisch-psychologische Diagnostik die methodische Grundlage und steuert theoretische Grundlagen für die Diagnose kognitiver Fähigkeiten, zu Fragen der Testkonstruktion sowie zur Datenanalyse bei.

Die Frage, ob eine inhaltlich relevante Differenz zwischen pädagogischer und pädagogisch-psychologischer Diagnostik konstatiert werden kann, wird in der aktuellen Literatur mehrheitlich verneint (vgl. bspw. Jäger, 2003, S. 313; Leutner, 2001, S. 521; Ingenkamp & Lissmann, 2008, S. 19). Vielmehr scheint die begriffliche Unterscheidung durch die Verwendung in verschiedenen Wissenschaftsdisziplinen etabliert worden zu sein: Während im Rahmen der Erziehungswissenschaft zumeist von pädagogischer Diagnostik die Rede ist, bevorzugt die (Pädagogische) Psychologie die Begrifflichkeit der pädagogisch-psychologischen Diagnostik. Leutner (2001) plädiert dafür, beide Begriffe synonym zu verwenden.

Traditionell relevante Kontexte der pädagogisch-psychologischen Diagnostik sind ursprünglich Ein- und Umschulungsentscheidungen, der Übertritt in die Sekundarstufe I und in den tertiären Bildungsbereich sowie Fragen der Berufsberatung und Berufsausbildung. Die häufigsten interventionsorientierten Anwendungsfelder sind die Diagnostik von Lern- und Verhaltensstörungen (vgl. bspw. Langfeldt & Tent, 1999) sowie die Hochbegabungsdiagnostik (vgl. bspw. Rost & Buch, 2010). So verstanden ist pädagogisch-psychologische Diagnostik eigentlich stets *individuumszentriert* und dient der Analyse interessierender Merkmale bei einzelnen Merkmalsträgern. Auch Aussagen, die sich auf Gruppen von Personen beziehen, basieren auf der Erhebung individueller Ist-Zustände, anhand derer die Personen homogenen Klassen von Merkmalsträgern zugeordnet werden können (Tent & Stelzl, 1993).

In den 70er und frühen 80er Jahren des letzten Jahrhunderts wurde an der pädagogischen Diagnostik teils heftige Kritik geübt. Beklagt wurden neben der einseitigen testpsychologischen Orientierung auch die mangelhaften theoretischen und methodologischen Grundlagen sowie die rationale Strategie der Entscheidungsfindung insgesamt. Im Zentrum der Kritik stand die vermeintliche Selektionsfunktion der pädagogisch-psychologischen Diagnostik. Als erstrebenswerte Alternative galt die Förderdiagnostik, da hierbei nicht mehr die Zuweisung der Schülerinnen und Schüler zu den einzelnen Schularten und damit die Interessen der Institution Schule zentral waren, sondern die optimale Förderung jedes Einzelnen im Fokus der Aufmerksamkeit stand. Auch heutzutage ist die Debatte um Selektions- und Förderdiagnostik noch nicht beigelegt. Ingenkamp und Lissmann (2008, S. 38) betonen, dass die Diskussion um diese beiden Strategien durch Missverständnisse und ideologische Argumente stark belastet ist. Sie kommen zu dem Schluss, dass die Förderung der individuellen Entwicklung ein zentrales pädagogisches Ziel darstellt, Selektion und Platzierung aber ebenfalls wichtig sind. In der Pädagogischen Psychologie haben daher – je nach Zielstellung – beide Strategien ihre Berechtigung.

Die Frage, ob, und wenn ja, inwieweit das Individuum der Bezugspunkt der pädagogisch-psychologischen Diagnostik ist oder ob auch andere Zielgrößen wie schulische Institutionen oder sogar das Bildungssystem als Ganzes in den Blick genommen werden können und sollen, ist nach wie vor Gegenstand der Diskussion. Neben dem oben dargestellten klassischen Konzept kann man daher auch einem weiter gefassten Verständnis von pädagogisch-psychologischer Diagnostik folgen, wie dies beispielsweise in der Definition von Klauer (1982) zum Ausdruck kommt: „Pädagogische Diagnostik ist das Insgesamt von Erkenntnisbemühungen im Dienste aktueller pädagogischer Entscheidungen“ (S. 5). Hier wird der Versuch deutlich, die pädagogisch-psychologische Diagnostik nicht über ihre Gegenstände und Aufgabenfelder zu definieren, da diese einem Wandel unterworfen und nicht abschließend bestimmbar sind (vgl. Klauer, 1982).

Einem so weit gefassten Begriff der pädagogisch-psychologischen Diagnostik könnte auch das Large-Scale-Assessment der empirischen Bildungsforschung zugeordnet werden. Ingenkamp und Lissmann (2008) betonen hingegen, dass Evaluation – wie zum Beispiel das Monitoring des deutschen Bildungssystems – ursprünglich nicht Aufgabe der pädagogischen Diagnostik sei. Diese Einschätzung basiert auf der Tatsache, dass bildungsbezogene Evaluationsmaßnahmen zwar Merkmale bei einzelnen Schülerinnen und Schülern erfassen, die persönliche Identität dieser Individuen aber nicht im Zentrum des Interesses steht. So sind zwar die Messvorgänge identisch, die Zielstellungen unterscheiden sich aber, da auf der einen Seite allgemeine Aussagen und auf der anderen Seite Aussagen im Hinblick auf individuelle Schülerinnen und Schüler und die Optimierung der für sie spezifischen pädagogischen Bedingungen angestrebt werden (vgl. Ingenkamp & Lissmann, 2008, S. 14).

Das Problem, dass Schulleistungsstudien zwar Erkenntnisse anstreben, die sich primär auf die Ebene des Schulsystems und nicht auf die Individualebene beziehen, aber dennoch pädagogisch diagnostisch relevant sind, da durch sie Voraussetzungen, Verlauf und Folgen individuellen Lernens erforscht werden, wird von Ingenkamp und Lissmann (2008, S. 14) allerdings selbst aufgeworfen und diskutiert. Sie ziehen den Schluss, dass ein erweitertes Verständnis der pädagogischen Diagnostik den veränderten Erkenntnisbedürfnissen entspricht und somit die Möglichkeit bietet, den drohenden Bedeutungsverlust der pädagogischen Diagnostik abzuwenden (vgl. Jäger, Frey, Wosnitza & Flor, 2001). Somit erscheint ihre Formulierung: „Diese evaluative Fragestellung wird zu einer Aufgabe der pädagogischen Diagnostik, weil wichtige Lernvoraussetzungen aufgearbeitet werden, die individuelles Lernverhalten beeinflussen und pädagogische Entscheidungen nach sich ziehen“ (Ingenkamp & Lissmann, 2008, S. 323) als Kompromiss zwischen dem traditionellen Selbstverständnis der pädagogisch-psychologischen Diagnostik und aktuell relevanten Erkenntnisbedürfnissen.

Ähnlich wie in der Erziehungswissenschaft (vgl. Abschnitt 2.1.3) wird auch aus der Perspektive der pädagogisch-psychologischen Diagnostik die zunehmende Orientierung an Testergebnissen und die steigende Zahl an Testdurchführungen im schulischen Kontext mitunter kritisch eingeschätzt. Insbesondere mangelhafte Rückmeldepraktiken und die unzureichende Nutzbarmachung von Testergebnissen für die unterrichtliche Arbeit der Lehrkräfte werden von Vertretern der Disziplin nachdrücklich beklagt: „Verordnete Tests, von denen die Wissenschaftler nach Genehmigung der Ministerien nur Mittelwerte und Verteilungen sowie Zusammenhänge mit Oberflächenvariablen mitteilen, helfen einer richtig verstandenen Pädagogischen Diagnostik nur wenig und werden Testaversionen wieder aufleben lassen“ (Ingenkamp & Lissmann, 2008, S. 9).

Durch ein steigendes Problembewusstsein gegenüber den Defiziten des Bildungssystems und durch das zunehmende Interesse an Schulleistungsdiagnostik ist man innerhalb der Pädagogischen Psychologie und der pädagogisch-psychologischen Diagnostik darum bemüht, durch theoretische Überlegungen und methodische Innovationen einen Beitrag zu den aktuellen Herausforderungen der empirischen Bildungsforschung zu leisten.

Neuere methodische Entwicklungen innerhalb der pädagogisch-psychologischen Diagnostik betreffen beispielsweise den zunehmenden Einsatz technologiebasierter Verfahren. Die Verwendung moderner elektronischer Hilfsmittel eröffnet nicht nur in der Testentwicklung und Testevaluation, sondern auch und insbesondere in der Testdurchführung, -auswertung und -rückmeldung neue Perspektiven. Für den hier relevanten Einsatzbereich des Large-Scale-Assessments kann computerbasiertes Testen unter anderem die Grundlage einer adaptiven Itemauswahl bilden (vgl. bspw. Frey, 2007). Ein weiteres Forschungsthema in diesem Kontext betrifft die Frage, ob von einer Äquivalenz verschiedener relevanter Konstrukte in unterschiedlichen Testmedien auszugehen ist, welche eine wesentliche Voraussetzung für den Einsatz technologiebasierter Testverfahren darstellt (vgl. bspw. Schroeders & Wilhelm, in press). Neben der Itementwicklung und Itemdarbietung kann computergestützte Diagnostik auch als Basis für die Implementierung neuer Testkonzepte und die Exploration neuer Konstrukte dienen (vgl. bspw. Kyllonen, Walters & Kaufman, 2011).

Als Problem innerhalb der pädagogisch-psychologischen Diagnostik wird mitunter diskutiert, dass in den letzten Jahrzehnten zwar verschiedene neue und interessante theoretische Konzepte und Instrumente ausgearbeitet wurden, Praktiker diesen Neuerungen gegenüber aber oftmals skeptisch seien und nach wie vor auf einen eingeschränkten und teils veralteten Vorrat an Testinstrumenten zurückgreifen. Aus wissenschaftlicher Perspektive bleibt die Praxis somit hinter aktuellen Entwicklungen und Forschungsergebnissen zurück (vgl. Langfeldt & Tent, 1999).

Insgesamt betrachtet befindet sich auch die pädagogisch-psychologische Diagnostik in einer Phase der partiellen Neuorientierung. Neben die traditionell etablierte individuumsbezogene Diagnostik treten zunehmend bildungsbezogene Evaluationsmaßnahmen. Hierbei wechselt zwar der Fokus der Diagnose, die methodischen Grundlagen und Messvorgänge bleiben jedoch weitgehend unverändert. Somit besteht der Beitrag der pädagogisch-psychologischen Diagnostik beziehungsweise der Pädagogischen Psychologie zum Forschungsfeld der empirischen Bildungsforschung primär in der Entwicklung und Bereitstellung von testmethodisch fundierten, aber auch innovativen Erhebungs- und Auswertungsverfahren (vgl. Prenzel, 2006).

2.1.3 Die empirische Bildungsforschung und ihr Verhältnis zur Erziehungswissenschaft

Die Erziehungswissenschaft setzt sich als wissenschaftliche Disziplin traditionell mit der Theorie und Praxis von Bildung und Erziehung auseinander. Innerhalb der Disziplin besteht allerdings kein vollständiger Konsens hinsichtlich ihrer inhaltlichen und forschungsmethodischen Ausrichtung. Vereinfacht gesprochen, stehen sich philosophisch-historisch orientierte, hermeneutisch arbeitende Vertreter einer klassischen Pädagogik und Vertreter einer analytischen, zumeist quantitativ-empirisch arbeitenden Erziehungswissenschaft gegenüber (vgl. bspw. Heid, 1996; Sailer, 2007). Die wissenschaftliche Orientierung letzterer wird mitunter auch als empirische Schulforschung (Ruhloff, 2007), oder – vorrangig von ihren Vertretern selbst – als empirische Bildungsforschung (vgl. bspw. Klieme, 2007; Prenzel, 2006) bezeichnet.

Die Divergenz zwischen traditioneller Theoriebildung und empirischer Forschung hat sich als Konsequenz der *Realistischen Wende* innerhalb der Erziehungswissenschaft in den frühen 60er Jahren des 20. Jahrhunderts verschärft. Nach Heid (1996) handelt es sich bei den skizzierten Positionen um nach wie vor unversöhnliche Gegensätze, die in der Frage nach dem Wissenschaftscharakter der Erziehungswissenschaft münden. Während in der Literatur einerseits geäußert wird, dass die beiden dargestellten Positionen wissenschaftstheoretisch sehr weit voneinander entfernt liegen (vgl. bspw. Sailer, 2007), wird eine solche Kontrastierung andernorts abgelehnt. Dass geisteswissenschaftliches Verstehen und naturwissenschaftliches Erklären ein unvereinbares Gegensatzpaar bilden, wurde bereits von Roth (1969) verneint: „Wir wünschen nicht nur zu verstehen, sondern auch zu erklären, und wer um Erklärung bemüht ist, sucht auch zu verstehen“ (S. 29). Dennoch scheint die Meinung zu überwiegen, dass Bildungstheorie und Bildungsforschung bis in die Gegenwart kaum vernetzt sind (vgl. Wigger, 2004). Ebenso konstatieren Tippelt und Schmidt (2010b), dass die Annäherung und Integration von empirischer Bildungsforschung und Allgemeiner Erziehungswissenschaft bislang nur bedingt gelungen sind (vgl. hierzu ausführlicher Abschnitt 4.5).

Auch eine mögliche Diskrepanz zwischen einem klassischen pädagogischen Bildungsbegriff und jenem der empirischen Bildungsforschung ist in der Literatur Gegenstand der Diskussion (vgl. bspw. Merkens, 2006). Tatsächlich ist es schwierig, *den* Bildungsbegriff der Pädagogik definieren zu wollen, da dieser in der Literatur facettenreich und vielschichtig beschrieben wird, ohne dass hierbei eine klare Konvergenz der Meinungen sichtbar werden würde. Bemüht man sich dennoch, den gemeinsamen Kern verschiedener Sichtweisen auf den pädagogischen Bildungsbegriff zu erfassen, so könnte dieser meinem Verständnis nach wie folgt formuliert werden: Der pädagogisch tradierte, auf den Vorstellungen der neuhumanistischen

Bewegung des 18. Jahrhunderts basierende Bildungsbegriff zielt auf Ganzheitlichkeit und betont die Eigenständigkeit des Individuums bei seiner Entwicklung. Nicht die Akkumulation von Wissen wird hier als Bildung verstanden, vielmehr stehen die Vorstellung des „vollendeten" und mündigen Menschen, der sich um die Ausbildung all seiner Fähigkeiten bemüht und die Entwicklung seiner Persönlichkeit im Zentrum des pädagogischen Interesses (vgl. bspw. Gloger-Tippelt, 2010; Merkens, 2006). Dass ein solches Begriffsverständnis aufgrund mangelnder Präzision problematisch sein kann, wird von Seiten der Erziehungswissenschaft – mitunter – selbst eingeräumt (vgl. Merkens, 2006). Alternativ zu einem pädagogisch geprägten Verständnis des Bildungsbegriffs etablieren sich auch psychologisch orientierte Sichtweisen (vgl. Götz, Frenzel & Pekrun, 2010). Auf der Ebene der theoretischen Begriffsbildung unterscheidet ein psychologisch orientiertes Verständnis des Bildungsbegriffs zwischen Bildung als Prozess und Bildung als Produkt. Hierbei umfasst die Prozessperspektive die Entwicklung und Vermittlung erwünschter Persönlichkeitsausprägungen, beispielsweise durch schulischen Unterricht. Die Produktperspektive zielt auf überdauernde Ausprägungen der Persönlichkeit eines Menschen, die pädagogisch erwünscht und in diesem Sinne normativ sind. Beispielsweise können Wissensbestände ebenso wie Werthaltungen und Verhaltensdispositionen als Produkte von Bildung verstanden werden (vgl. Götz, Frenzel & Pekrun, 2010).

In der empirischen Bildungsforschung vorherrschend und auch innerhalb der bildungspolitischen Diskussion häufig gebraucht ist ein pragmatisch orientierter, *kompetenzbasierter Bildungsbegriff* (vgl. Abschnitt 2.3). Ein solches Begriffsverständnis verweist auf das Ziel, Bildung innerhalb des Systems und für jeden Einzelnen effektiver, nachhaltiger und ergebnisorientierter zu gestalten. Diese funktional-pragmatische Sichtweise lehnt sich an das angloamerikanische *Literacy*-Konzept an, welches auf einer Alltags- und Anwendungsorientierung grundlegender Kulturtechniken basiert (vgl. Zeitlinger, 2009, S. 55).

Ein solches kompetenzorientiertes Begriffsverständnis findet auch zunehmend in der Bildungspolitik und -administration Anwendung, so beispielsweise im 12. Kinder- und Jugendbericht, der sich schwerpunktmäßig mit Fragen der Bildung, Betreuung und Erziehung vor und neben der Schule auseinandersetzt (Bundesministerium für Familie, Senioren, Frauen und Jugend, [BMFSFJ], 2005).

Einigen Vertretern der traditionellen Pädagogik sowie der Allgemeinen Erziehungswissenschaft erscheint der klassische Bildungsbegriff mit dem Selbstverständnis der empirischen Bildungsforschung und standardisierten Leistungsmessung als generell inkompatibel (vgl. bspw. Gruschka, 2006a, 2006b, 2007; Merkens, 2006; Ruhloff, 2007).

Baumert (2001) äußert in diesem Zusammenhang, dass sich die wahrgenommene Unvereinbarkeit des pädagogischen Bildungsbegriffs mit standardisierter Leistungsmessung und

insofern mit der empirischen Bildungsforschung auf zwei zentrale Einwände zurückführen lässt. Zum einen ist dies der (behauptete) Widerspruch zwischen der Ganzheitlichkeit von Bildungsprozessen und den eingeschränkten Fragestellungen von Evaluationsmaßnahmen, die auf standardisierte Leistungsmessung zurückgreifen. Zum anderen werde der Bildungsbegriff, welcher der Messung zugrunde liegt, nicht hinreichend theoretisch untermauert, was dazu führe, dass Bildungsqualität mit dem gleichgesetzt werde, was der Test misst (Baumert, 2001, S. 24f.). Tatsächlich trifft es zu, dass Leistungstests nicht das gesamte Spektrum schulischer Ziele abbilden, sondern notwendigerweise eine Auswahl treffen. Diese Auswahl bezieht sich jedoch auf unentbehrliche Kernziele schulischer Bildung, wie die Beherrschung der Verkehrssprache in Wort und Schrift oder das Mathematisieren. Diese Aspekte stellen für die gesellschaftliche Teilhabe, das Berufsleben und die pädagogisch angestrebte Mündigkeit und Selbstentfaltung unerlässliche Voraussetzungen dar (vgl. bspw. Köller, 2008). Baumert (2001) weist hinsichtlich der Selektivität von Testinhalten ergänzend auf Folgendes hin: „In der Regel sind diese Präferenzen – was häufig übersehen wird – bereits aus den Stundentafeln und Vorschriften über die Leistungsfeststellungen in der Schule zu ersehen, […]“ (S. 25f.).

Die hier skizzierte Diskussion lässt sich auf die länderübergreifenden Bildungsstandards übertragen (vgl. Abschnitt 2.2.2), denen von Vertretern der Pädagogik ebenfalls grundsätzliche Zweifel entgegengebracht werden (vgl. Gruschka, 2006a, 2006b, 2007). Im Zuge der Entwicklung und Einführung der Bildungsstandards und der zunehmenden Prominenz des Kompetenzgedankens ist von Vertretern der empirischen Bildungsforschung jedoch wiederholt auf die Kompatibilität des Bildungsbegriffs im Sinne einer klassischen Bildungstheorie mit den Methoden der empirischen Bildungsforschung hingewiesen und für ein in diesem Sinne modernes Bildungsverständnis geworben worden (vgl. bspw. Klieme et al., 2007; Köller, 2008).

Von Vertretern der traditionellen Pädagogik wird das Bemühen um eine Annäherung hinsichtlich zentraler Inhalte und Begriffe jedoch teils vehement zurückgewiesen. Gruschka bemerkt unter Bezugnahme auf die so genannte „Klieme-Expertise“ (Klieme et al., 2007): „Während noch die theoretisch etwas zurückgebliebenen empirischen Bildungspolitiker wie Weiler oder so manche pädagogischen Psychologen nicht wissen, was sie mit diesem Nebulosum-Numinosum ‚Bildung’ anfangen können […], weil sie dessen wissenschaftlich gebotene Operationalisierbarkeit vermissen und sie schon deswegen wissenschaftlich zur Kompetenzforschung greifen müssen, die sie freilich (warum eigentlich?) ‚Bildungsforschung’ nennen, erklärt die Expertise: Alles nur ein Missverständnis! Es gibt eine unbemerkt gebliebene Konvergenz der Denkformen […]“ (Gruschka, 2006a, S. 10). Köller (2008, S. 50) bezeichnet die Äußerung, dass Bildungsstandards inkompatibel mit traditionellen Zielen und Ansprüchen der

Bildungstheorie wären, als Missverständnis und verweist ebenfalls auf die Relevanz einer kompetenten Nutzung von grundlegenden Kulturtechniken.

Nichtsdestotrotz ist es zutreffend, dass bestimmte schulische Erfahrungen, die mit dem klassischen Bildungsbegriff durchaus vereinbar sind, in der gegenwärtig von Bildungsstandards und Kompetenzorientierung geprägten Diskussion zunehmend aus dem Blickfeld geraten. Hierbei handelt es sich beispielsweise um ästhetische literarische Erfahrungen (vgl. Spinner, 2008, S. 321f.). Ferner kann durch die aktuelle Fokussierung auf Kompetenzen und Standards sowie die hiermit verbundene Output-Orientierung der Eindruck entstehen, „dass Lernergebnisse Produkte seien, auf die man nur genau und richtig hinarbeiten müsse“ (Zeitlinger, 2009, S. 54). Lernen hat jedoch deutlich mehr Prozess- als Produktcharakter, wobei der Lernprozess oft sehr komplex und stets multifaktoriell bedingt ist. Beschränkt man sich daher auf die Betrachtung der zu erzielenden Lernergebnisse und konzipiert diese ausschließlich als Produkt, ohne dem dahinterstehenden Prozess die erforderliche Aufmerksamkeit zu schenken, so könnte dies tatsächlich als bildungstheoretischer Reduktionismus gedeutet werden (vgl. Zeitlinger, 2009; Ziener, 2008). Die empirische Bildungsforschung beschäftigt sich jedoch nicht ausschließlich – und eventuell nicht einmal primär – mit Schulleistungstests. Ebenso relevant ist die quantitative wie qualitative empirische Erforschung unterrichtlicher Lehr-Lernprozesse mit dem Ziel, fach- und methodenspezifische wie -übergreifende Merkmale guten Unterrichts und gelingenden Lernens zu bestimmen (vgl. bspw. Helmke et al., 2007). „Empirische Schulforschung produziert eben nicht ‚chaotisch flottierende Daten’, die erst nachträglich beziehungsweise von außen, z. B. durch die Allgemeine Erziehungswissenschaft ‚in den zugehörigen pädagogischen Sachzusammenhang’ eingebettet werden müssen […]“ (Klieme, 2007, S. 144, als Erwiderung auf Ruhloff, 2007). Vielmehr ist empirische Schul- und Bildungsforschung substantielle, theoriegeleitete Forschung hinsichtlich pädagogischer Sachverhalte, die gesellschaftlich relevante Ergebnisse zur Verfügung stellt (vgl. Klieme, 2007).

In dem erläuterten Sinne wird von Seiten der traditionellen Pädagogik auch der Begriff der *Bildungsforschung* bisweilen in Frage gestellt (vgl. Gruschka, 2006a). Dieser fand aber bereits in den 70er Jahren des 20. Jahrhunderts häufig Verwendung. Der Deutsche Bildungsrat (1974) bestimmt den Gegenstand der Bildungsforschung als Untersuchung der Voraussetzungen und Möglichkeiten von Bildungs- und Erziehungsprozessen im institutionellen und gesellschaftlichen Kontext (Deutscher Bildungsrat 1974, S. 16). Bereits damals etablierte der Deutsche Bildungsrat in den Empfehlungen der Bildungskommission eine Unterscheidung von zwei Bedeutungsnuancen des Begriffs Bildungsforschung: „Man kann Bildungsforschung im weiteren und engeren Sinne auslegen. Im engeren Sinne hat es sie als Unterrichtsforschung immer schon gegeben. Im weiteren Sinne kann sie sich auf das gesamte Bildungswesen und seine Reform im

Kontext von Staat und Gesellschaft beziehen, einschließlich der außerschulischen Bildungsprozesse." (Deutscher Bildungsrat, 1974, S. 16). Das aktuelle öffentliche Interesse und der in den letzten zwei Jahrzehnten stattfindende, weitreichende Ausbau der Bildungsforschung betreffen vorrangig das weit gefasste Verständnis und belegen die wachsende gesellschaftliche Aufmerksamkeit für Bildungsthemen.

Verschiedenste Problemlagen im Bildungsbereich verlangen nach empirisch abgesicherten Erkenntnissen und der Entwicklung von Handlungskonzepten (vgl. bspw. König & Zedler, 2004; Prenzel, 2006). Von Relevanz sind hier unter anderem Fragen der Organisation und Gestaltung vorschulischer Bildung, Betreuung und Erziehung (vgl. bspw. Zwölfter Kinder- und Jugendbericht, BMFSFJ, 2005; Gloger-Tippelt, 2010) oder der institutionellen Strukturen schulischer Bildung im Hinblick auf die Mehrgliedrigkeit des Schulsystems (vgl. bspw. Dedering & Holtappels, 2010). In der aktuellen wissenschaftlichen Diskussion wird geäußert, dass die traditionelle Pädagogik die gesellschaftlich relevanten Fragen nicht zufriedenstellend beantworten und auch „die Erwartungen der praktisch tätigen Pädagogen auf konkretes Handlungswissen nicht erfüllen" konnte (König & Zedler, 2004, S. 78; vgl. auch Hansel, 2007, S. 190). Diesen Gedanken aufgreifend führt auch Prenzel (2006, S. 77) die gegenwärtige Relevanz der empirischen Bildungsforschung auf die akute gesellschaftliche Bedeutung des Gegenstandsbereiches zurück. Dass sich dies wechselseitig bedingt und die Relevanz bildungsbezogener Themen in Politik und öffentlicher Diskussion auch aufgrund empirischer Bildungsforschung und ihrer Ergebnisse ins Zentrum der Aufmerksamkeit rückten, würdigt Hansel (2007): „Es ist das Verdienst u.a. der empirischen Bildungsforschung, dass sie – geradezu über Nacht – diese Themenpalette zurück in einen öffentlichen Focus gerückt und die Erziehungswissenschaft erneut auf den Plan gerufen hat" (S. 195f.). Die Ergebnisse empirischer Bildungsforschung können auf verschiedene Weise – direkt und indirekt – gesellschaftlich handlungswirksam werden und dies nicht nur im Sinne einer Maßnahmenforschung, sondern auch im Sinne einer Orientierungsforschung, die empirisch gewonnene Hinweise für Handlungsstrategien und Reformansätze bietet (vgl. Tippelt & Schmidt, 2010b). Empirische Bildungsforschung liefert die Datengrundlage und damit die Voraussetzung für wissensbasierte, rationale Entscheidungen von pädagogischen Innovationen und Reformprozessen und erfüllt in diesem Sinne gesellschaftlich an sie herangetragene Orientierungs-, Aufklärungs- und Steuerungsfunktionen. Dass auch dieser Aspekt von einigen Vertretern der Erziehungswissenschaft äußerst kritisch bewertet wird (vgl. bspw. Ruhloff, 2007), wird in Abschnitt 4.5 dieser Dissertation eingehender diskutiert.

2.1.4 Einbettung der vorliegenden Arbeit in die vorgestellten Disziplinen

Diese Dissertation ist durch ein stark interdisziplinäres Verständnis empirischer Bildungsforschung geprägt. Dies entspricht der Darstellung in aktuellen Publikationen (vgl. Prenzel, 2006; Tippelt & Schmidt, 2010a), welche die empirische Bildungsforschung als ein Forschungsgebiet mit inter- und multidisziplinärem Charakter kennzeichnen, an dem unter anderem die Pädagogische Psychologie, die Entwicklungspsychologie, die Fachdidaktiken, die Erziehungswissenschaft sowie die Bildungssoziologie und -ökonomie beteiligt sind (vgl. Tippelt & Schmidt, 2010b, S. 9f.).

Diese Arbeit leistet einen Beitrag zur empirischen Bildungsforschung, indem sie insbesondere Stärken der pädagogisch-psychologischen Diagnostik mit denen der Fachdidaktik Deutsch vereint. Verschiedene Vertreter von Erziehungswissenschaft und empirischer Bildungsforschung betonen, dass es der Psychologie gelungen ist, sehr hohe methodische Standards zu etablieren (vgl. Prenzel, 2006, S. 76). Nimmt man dies als gegeben, scheint der Versuch lohnenswert, Vorsprünge im Hinblick auf methodisches Handwerkzeug in den Dienst inhaltlich reicher und erziehungswissenschaftlich praktisch nutzbarer Forschung zu stellen. Methodisch exzellentes Arbeiten kann nämlich nur dann sinnvolle Ergebnisse generieren, wenn diese Methoden inhaltlich sinnvoll zum Einsatz kommen (vgl. Groeben, 2005). Daher möchte die vorliegende Arbeit dem Aufruf von Bremerich-Vos (2002) folgen und sich bemühen, den Bedarf an erziehungswissenschaftlicher wie fachdidaktischer Fundierung empirischer Bildungsforschung einzulösen. Konkret umfasst dies in den nachfolgenden empirischen Beiträgen die theoretische Herleitung der Konstruktdefinitionen und die Testentwicklung, insbesondere in den Bereichen Sprachgebrauch (Abschnitt 3.5), Schreiben (Abschnitte 3.3.1 sowie II) und Rechtschreibung (Abschnitte 3.3.2 sowie III). So wären beispielsweise die in dem Beitrag zur Diagnostik der Rechtschreibkompetenz in der Grundschule (Böhme & Bremerich-Vos, 2009) vorgenommenen Strukturprüfungen ohne die deutschdidaktische Fundierung der qualitativen Fehleranalysen nicht möglich gewesen. In anderen Bereichen wurde neben der Einbeziehung der deutschdidaktischen Literatur auch auf andere Theoriegrundlagen zurückgegriffen. Dies betrifft zum Beispiel den Bereich des Leseverstehens, da es hier eine prominente und ausgereifte lesepsychologische Forschung gibt, die auch von Seiten der Deutschdidaktik rezipiert wird (vgl. bspw. Christmann, 2010). Ferner betrifft dies den Bereich der Zuhörforschung, der auch innerhalb der Deutschdidaktik ein sehr junges Forschungsfeld darstellt und ebenfalls psychologisch geprägt ist (vgl. Imhof, 2003, 2010).

Soll Bildungsforschung in dem Sinne erfolgreich sein, dass sie von der Gesellschaft nachgefragte Sachverhalte bearbeitet und zur Klärung von relevanten Fragestellungen beiträgt,

dann scheint eine interdisziplinäre Kooperation unerlässlich. Diesem Verständnis folgend, möchte die hier vorliegende Arbeit einen wissenschaftlichen Beitrag zur empirischen Bildungsforschung leisten, indem sie aktuelle, politisch initiierte Bemühungen zur empirischen Untersuchung schulischer Bildungserträge mit Hilfe der Etablierung und Überprüfung länderübergreifender Bildungsstandards analysiert. Diese Untersuchungen sind aber stets in einen deutschdidaktischen Kontext eingebettet, der sowohl als Ausgangspunkt der Diagnostik als auch für die Interpretation der Ergebnisse fungiert und damit auch die pädagogische Nutzbarmachung empirischer Befunde im Blick behält. Diese Perspektive ist sowohl für die Rahmung als auch für die empirischen Beiträge leitend.

Abschließend möchte ich mit Klieme und Kollegen festhalten: „Das Zusammenstellen und Erproben von Tests und schließlich der Einsatz von Tests im Rahmen der schulübergreifenden Qualitätssicherung und -entwicklung sind jedoch weitergehende, spezialisierte Tätigkeiten, die ein Zusammenspiel von Experten aus der Fachdidaktik, der empirischen Bildungsforschung und der pädagogisch-psychologischen Methodenlehre erfordern“ (Klieme et al., 2007, S. 81).

2.2 Die länderübergreifenden Bildungsstandards in Deutschland

2.2.1 Hintergründe der Einführung länderübergreifender Bildungsstandards

Die Ständige Konferenz der Kultusminister der Länder der Bundesrepublik Deutschland (Kultusministerkonferenz, KMK) leitete zu Beginn des neuen Jahrtausends eine grundlegende Umorientierung der deutschen Bildungspolitik ein. Diese zeichnet sich durch die Umstellung von einer Input- hin zu einer Outputsteuerung aus, welche damit einhergeht, dass nicht länger Lerninhalte durch Lehrpläne oder Curricula definiert werden, sondern Bildungsergebnisse die zentrale Größe in der Steuerung des deutschen Bildungssystems darstellen. Koeppen, Hartig, Klieme und Leutner (2008) bestimmen Bildungsergebnisse wie folgt: „The outcomes of education are the knowledge acquired, the abilities, skills, attitudes, and dispositions developed, and the qualifications attained“ (S. 61). In der gegenwärtigen Diskussion findet in diesem Zusammenhang zumeist der Begriff der schulisch erworbenen *Kompetenzen* Anwendung (zur Diskussion des Kompetenzbegriffs vgl. Abschnitt 2.3.). Dass die angestrebten Lernergebnisse von den Schülerinnen und Schülern tatsächlich erreicht werden, liegt auch in der Verantwortung der Bildungsadministration sowie der jeweiligen Schulen (vgl. Klieme, et al., 2007). Insgesamt wird nun verstärkt das Ziel verfolgt, das Bildungssystem sowie die schulische Bildung jedes

Einzelnen in Anlehnung an das angloamerikanische *Literacy*-Konzept nachhaltiger und ergebnisorientierter zu gestalten (vgl. Zeitlinger, 2009, S. 55) und die Verfügbarkeit grundlegender Kulturtechniken sicherzustellen.

Ausgangspunkt dieser Umorientierung waren die erwartungswidrig schwachen deutschen Ergebnisse in den internationalen Vergleichsstudien TIMSS II/III im Jahr 1995 (Baumert et al., 1997; Baumert, Bos & Lehmann, 2000a) und PISA 2000 (Baumert et al., 2001; Baumert et al., 2002). Bis auf wenige Ausnahmen waren diese Studien die ersten auf standardisierten Leistungstests basierenden Überprüfungen deutscher Bildungsergebnisse im internationalen Vergleich. Da in Deutschland bis zum Beginn der 1990er Jahre kaum belastbare empirische Daten zu den Erträgen schulischer Bildung vorlagen, wurde den Ergebnissen der internationalen Vergleichsstudien – insbesondere aufgrund ihrer für Deutschland unerfreulichen Resultate – in Politik, Wissenschaft und Öffentlichkeit besondere Beachtung geschenkt. Köller (2008) spricht in diesem Zusammenhang davon, dass die in TIMSS berichteten Leistungsstände, nach denen 1995 etwa 50% der deutschen Schülerinnen und Schüler am Ende der Pflichtschulzeit den Kernzielen mathematischer Grundbildung nicht gerecht werden konnten (vgl. Baumert, Bos & Lehman 2000b), das Bildungssystem „in tiefe Nachdenklichkeit" stürzten (Köller, 2008, S. 49). Neben den im internationalen Vergleich enttäuschenden mittleren Leistungsständen deutscher Schülerinnen und Schüler verweisen die zitierten Studien auf weitere entscheidende Defizite des deutschen Bildungssystems, die sich zum Beispiel in besonders stark ausgeprägten Disparitäten manifestieren. Diese Disparitäten betreffen unter anderem Leistungsdifferenzen zwischen verschiedenen Regionen Deutschlands, zwischen Kindern aus unterschiedlichen sozialen Schichten sowie zwischen Kindern mit und ohne Migrationshintergrund (vgl. bspw. Klieme et al., 2007).

Die in den internationalen Vergleichsstudien aufgedeckten Defizite haben somit wiederholt deutlich gemacht, dass in Deutschland für einen Anschluss an die internationale Leistungsspitze umfassende Bildungsreformen erforderlich sind und die Qualität des deutschen Bildungssystems nachhaltig verbessert und langfristig stabilisiert werden muss. „Das klassische Vertrauen darauf, dass das, was gelehrt wird, auch gelernt wird, hat sich als nicht haltbar erwiesen" (Merkens, 2006, S. 17).

Als erste Reaktion auf das unerwartet schwache Abschneiden der deutschen Jugendlichen definierte die Kultusministerkonferenz die Qualitätssicherung im deutschen Bildungswesen bereits im Oktober 1997 in den „Konstanzer Beschlüssen" als zentrale Zielstellung. Diese Ausgangssituation wird in der Literatur zumeist als Nährboden der empirischen Wende in der Erziehungswissenschaft charakterisiert, in deren Folge die empirische Bildungsforschung sowie

die pädagogisch-psychologische Diagnostik innerhalb dieser Wissenschaftsdisziplin deutlich an Gewicht gewannen (vgl. Abschnitt 2.1.3).

Ein Erfolg versprechender Schritt auf dem Weg zur Sicherstellung der Qualität des deutschen Bildungssystems ist eine differenzierte Analyse der Bildungs- und Schulsysteme und somit der Erfolgsfaktoren jener Staaten, die in internationalen Vergleichsstudien wiederholt besonders gute Ergebnisse erzielten. Mit dieser Intention wurde vom Bundesministerium für Bildung und Forschung eine Studie in Auftrag gegeben, in der durch systematische Vergleiche verschiedener Bildungssysteme solche Reformansätze und Steuerungsmodelle identifiziert wurden, die den in Schulleistungsstudien erfolgreicheren Staaten gemeinsam sind (vgl. für einen Überblick Böhme, 2006; Arbeitsgruppe „Internationale Vergleichsstudie", 2007; van Ackeren, 2007). Als wesentliche Faktoren wurden neben einer Ausweitung der Eigenverantwortung der Schulen, einer eher spät einsetzenden Differenzierung in verschiedene Bildungsgänge und einer intensiven individuellen Förderung der Schülerinnen und Schüler vor allem die Etablierung länderübergreifender Bildungsstandards sowie die regelmäßige, professionelle Durchführung von zentralen Vergleichsstudien identifiziert (Arbeitsgruppe „Internationale Vergleichsstudie", 2007; van Ackeren, 2007).

Zu Beginn des neuen Jahrtausends wurden im Auftrag der Kultusministerkonferenz länderübergreifende Bildungsstandards für verschiedene Fächer und unterschiedliche Schulabschlüsse entwickelt. Mit diesem Schritt wurden wichtige Erfahrungen anderer Staaten nutzbringend für eine positive Entwicklung des deutschen Bildungssystems umgesetzt. Tabelle 1 auf der folgenden Seite gibt einen Überblick zu den gegenwärtigen Entwicklungsständen der Bildungsstandards für verschiedene Fächer und Abschlüsse.

In den Jahren 2003 und 2004 wurden die länderübergreifenden Bildungsstandards für die Primarstufe und die Sekundarstufe I von der Kultusministerkonferenz als verbindlich verabschiedet. Parallel wurde die Einrichtung des *Instituts zur Qualitätsentwicklung im Bildungswesen (IQB)* angestoßen, welches die Aufgabe hat, die länderübergreifenden Bildungsstandards weiterzuentwickeln, zu operationalisieren, zu normieren und ihre Erreichung zu überprüfen. Die empirische Überprüfung der Bildungsstandards ist in eine Gesamtstrategie der Kultusministerkonferenz für das Monitoring des deutschen Bildungssystems integriert (KMK, 2006). Diese Strategie umfasst vier Säulen:

- die stichprobenbasierte Beteiligung an internationalen Schulleistungsstudien: PISA in der Sekundarstufe I sowie PIRLS und TIMSS in der Primarstufe,
- die stichprobenbasierte Überprüfung der Erreichung der länderübergreifenden Bildungsstandards in der Primarstufe sowie in der Sekundarstufe I in

Ländervergleichsstudien, die in Ankopplung an die internationalen Large-Scale-Assessments durchgeführt werden,

- die regelmäßige Realisierung von flächendeckenden Vergleichsarbeiten (VERA) in der dritten sowie achten Jahrgangsstufe, die sich ebenfalls auf die länderübergreifenden Bildungsstandards beziehen und Impulse für die kompetenzorientierte Unterrichtsentwicklung und Förderung der diagnostischen Kompetenz der Lehrkräfte bieten sollen sowie
- die gemeinsame Bildungsberichterstattung von Bund und Ländern.

Tabelle 1: Entwicklungsstand der Bildungsstandards für verschiedene Fächer und Abschlüsse

Fach	Primarstufe (Ende der 4. Jahrgangsstufe)	Sekundarstufe I		Sekundarstufe II (Allgemeine Hochschulreife)
		Hauptschul-abschluss	Mittlerer Schulabschluss	
Deutsch	2004 verabschiedet	2004 verabschiedet	2003 verabschiedet	2011 in Entwicklung
Mathematik	2004 verabschiedet	2004 verabschiedet	2003 verabschiedet	2011 in Entwicklung
Erste Fremdsprache (Englisch, Französisch)	---	2004 verabschiedet	2003 verabschiedet	2011 in Entwicklung
Naturwissenschaften (Biologie, Chemie, Physik)	---	---	2004 verabschiedet	---

Nach Einschätzung der Kultusministerkonferenz dient diese Gesamtstrategie zum Bildungsmonitoring der systematischen Beschaffung von Informationen über das deutsche Bildungssystem (KMK, 2006). Um positive Entwicklungsimpulse entfalten zu können, muss diese Informationsvielfalt jedoch unverzichtbar mit konkreten Maßnahmen zur Unterrichts- und Qualitätsentwicklung in den Schulen vor Ort verknüpft werden. Hierzu äußert die KMK (2006): „Damit die verschiedenen Maßnahmen die notwendigen Impulse zur Verbesserung des Bildungswesens auch tatsächlich auslösen, ist es erforderlich, Prozesse der Qualitätsentwicklung und Standardsicherung auf allen Ebenen, von der einzelnen Schule bis zum gesamten Bildungssystem, systematisch umzusetzen und miteinander zu verbinden. Insbesondere muss sichergestellt werden, dass Informationen über die Qualität des Bildungssystems so weit wie möglich auch für die Entwicklung jeder einzelnen Schule genutzt werden können“ (S. 6).

2.2.2 Charakterisierung der länderübergreifenden Bildungsstandards

„Bildungsstandards werden zunächst verbal formuliert. Sie benennen die Kompetenzen, die Schülerinnen und Schüler im jeweiligen Lernbereich erwerben sollen, und stützen sich dabei auf Kompetenzmodelle, in denen Teilaspekte (Dimensionen) und Stufen dieser Kompetenzen spezifiziert werden“ (Klieme et al., 2007, S. 81).

Für die Einführung der länderübergreifenden Bildungsstandards in Deutschland stellte die so genannte Klieme-Expertise (Klieme et al., 2007) eine wesentliche Grundlage dar. Sie wurde im Jahr 2002 im Auftrag des Bundesministeriums für Bildung und Forschung angefertigt und umfasst Stellungnahmen zur Konzeption und Funktion von Bildungsstandards, zu den Grundlagen ihrer Entwicklung, zu möglichen Konsequenzen der Einführung für das Bildungssystem sowie zur Implementation von Bildungsstandards in Deutschland. Die Expertise integriert die Perspektiven von Vertretern sehr unterschiedlicher Fachgebiete und vereint hierbei nicht nur die Sichtweisen der Allgemeinen Erziehungswissenschaft mit jener der empirischen Bildungsforschung, sondern bezieht ferner Fachdidaktiken, die pädagogisch-psychologische Methodik sowie bildungsrechtliche und historische Standpunkte mit ein (vgl. Klieme et al., 2007, S. 15). Aus Sicht der Klieme-Expertise benennen Bildungsstandards Ziele der pädagogischen Arbeit und konkretisieren den Bildungsauftrag, den allgemeinbildende Schulen erfüllen sollen (Klieme et al., 2007, S. 19).

Konkreter lassen sich die länderübergreifenden Bildungsstandards als verbindlich und fachspezifisch formulierte Kernziele schulischer Bildung charakterisieren. Sie orientieren sich an den Grundprinzipien des jeweiligen Unterrichtsfaches und beschreiben wesentliche fachbezogene Kompetenzen, die die Schülerinnen und Schüler bis zu einem bestimmten Zeitpunkt in ihrer Bildungsbiografie erreicht haben sollen. Die erwünschten beziehungsweise erwarteten Lernergebnisse formulieren die Standards mit Hilfe so genannter *Can-Do-Statements* und orientieren sich hierbei am Kompetenzkonzept (vgl. Abschnitt 2.3; Klieme et al., 2007; Weinert, 2001a). Mit der verbindlichen Einführung der länderübergreifenden Bildungsstandards stehen somit nicht länger die in Lehrplänen definierten Unterrichts*inhalte*, sondern die von den Schülerinnen und Schülern zu erwerbenden Kompetenzen im Fokus der Aufmerksamkeit. Hiermit verbindet sich die Hoffnung, dass sich auch die Arbeit der Lehrkräfte im Unterrichtsalltag stärker am Kompetenzaufbau ausrichtet.

Nach Klieme und Kollegen (2007, S. 24ff.) sowie Köller (2010, S. 530f.) zeichnen sich „gute“ Bildungsstandards neben ihrer Fachspezifität und Fokussierung auf die Kernbereiche des jeweiligen Faches durch die Förderung kumulativer Lernprozesse, eine verständliche Darstellung und realistische Zielstellungen aus. Ferner bieten sie die Möglichkeit der Differenzierung

verschiedener Leistungsniveaus je nach den schülerseitigen sowie im Unterrichtsprozess vorliegenden Lernvoraussetzungen. Ein weiteres erfolgskritisches Merkmal guter Bildungsstandards ist ihre Messbarkeit. Das heißt, dass die Standards so formuliert sind, dass sich Messinstrumente zu ihrer Überprüfung ableiten lassen (vgl. Köller, 2010, S. 531).

Den genannten Merkmalen können die Bildungsstandards je nach Fach unterschiedlich gut gerecht werden. Köller (2010) spricht davon, dass „die Standards im Fach Deutsch in Teilen die Kriterien Fokussierung, Verständlichkeit, Realisierbarkeit und Messbarkeit vermissen" (S. 534) lassen, während die Standards für das Fach Mathematik die überwiegende Mehrzahl der gewünschten Eigenschaften aufweisen. Insbesondere auf den Umstand, dass sich verschiedene der für das Fach Deutsch formulierten Standards nicht beziehungsweise nicht ohne Weiteres für eine standardisierte Leistungsüberprüfung anbieten, ist sowohl von Seiten der Fachdidaktik (vgl. bspw. Spinner, 2008) als auch von Seiten der empirischen Bildungsforschung (vgl. Granzer, Böhme & Köller, 2008) hingewiesen worden. Als Vertreter der Deutschdidaktik äußert sich Spinner in diesem Zusammenhang wie folgt: „Die Formulierung von Bildungsstandards, die Erarbeitung von Kompetenzmodellen und die Entwicklung von Testaufgaben sind im Fach Deutsch besonders schwierig" (Spinner, 2008, S. 313).

Trotz dieser Problematik stellen die Standards einen länderübergreifend verbindlichen Vergleichsmaßstab dar, der sowohl für interne als auch für externe Evaluationsprozesse wichtige Bewertungskriterien zur Verfügung stellt und auf diese Weise zur Verbesserung der Leistungsfähigkeit des Bildungssystems beitragen kann. Einhergehend mit der Einführung der Bildungsstandards gewinnen großflächig angelegte Leistungserhebungen mittels standardisierter Tests zunehmend an Bedeutung. Derartige Testungen dienen nicht nur der Überprüfung und Normierung der Bildungsstandards, sondern auch einem kontinuierlichen, länderübergreifenden Bildungsmonitoring. Schulleistungsstudien sollten also keinesfalls als bloße Bestandsaufnahme verstanden werden, vielmehr stellen die in ihnen gewonnenen empirischen Befunde eine essentielle Komponente der ergebnisorientierten Systemsteuerung und damit einen wesentlichen Ausgangspunkt der Qualitätsentwicklung und -sicherung im deutschen Bildungssystem dar (Böhme, 2006).

Köller (2008, S. 50) sieht die Innovationskraft der Bildungsstandards darin, dass sie auf Kompetenzen verweisen und damit eine Abkehr von der Input-Orientierung darstellen, bei der mittels Lehrplänen und Curricula eine starke Steuerung der Unterrichtsinhalte erfolgte. Dadurch, dass die länderübergreifenden Bildungsstandards der Output-Orientierung und dem Kompetenzkonzept verpflichtet sind, erhalten die Lehrkräfte größere Freiräume bei der Unterrichtsgestaltung, die in deutlich geringerem Umfang durch inhaltliche Vorgaben eingeengt wird.

Gelegentlich wird in der fachwissenschaftlichen Diskussion geäußert, dass die durch die Bildungsstandards und ihre Evaluierung prominente Output-Orientierung dem traditionellen Selbstverständnis der Erziehungswissenschaft und auch der Profession des Pädagogen entgegenstehe. Die gegenwärtige Standardorientierung führe zu negativen Tendenzen, die beispielsweise von Shirley als „historical amnesia, educational reductionism, and spiritual displacement" charakterisiert werden (Shirley, 2008, S. 35). Insbesondere der Aspekt, dass der breite schulische Bildungskanon zu Gunsten einer testvorbereitenden Vermittlung von Kernkompetenzen (Lesen, Schreiben und mathematische Grundbildung) leide, wird in der deutschen und internationalen Literatur kritisch diskutiert (vgl. hierzu Abschnitt 4.4).

Die Kritik, Bildungsstandards wären mit den traditionellen Zielen und Ansprüchen der Bildungstheorie nicht vereinbar, weist Köller (2008, 2010) zurück und betont, dass sich bereits die Klieme-Expertise (Klieme et al., 2007) um eine Auseinandersetzung mit der bildungstheoretischen Verankerung der Bildungsstandards bemüht habe. Klieme und Kollegen (2007) kamen hierbei zu der folgenden Einschätzung:

> „Kompetenzen" beschreiben aber nichts anderes, als solche Fähigkeiten der Subjekte, die auch der Bildungsbegriff gemeint und unterstellt hatte: Erworbene, also nicht von Natur aus gegebene Fähigkeiten, die an und in bestimmten Dimensionen der gesellschaftlichen Wirklichkeit erfahren wurden und zu ihrer Gestaltung geeignet sind, Fähigkeiten zudem, die der lebenslangen Kultivierung, Steigerung und Verfeinerung zugänglich sind, so, dass sie sich intern graduieren lassen, z. B. von der grundlegenden zur erweiterten Allgemeinbildung […]. (Klieme et al., 2007, S. 65)

Köllers (2008, 2010) zentrales Argument bezieht sich auf die Tatsache, dass die Standards die Kernelemente kultureller Literalität beschreiben und damit wesentliche Voraussetzungen für persönliches Wachstum und eine mündige gesellschaftliche Teilhabe darstellen. In diesem Sinne stehen die Bildungsstandards den klassischen humanistischen Bildungsidealen nicht entgegen, sondern thematisieren basale Bestandteile von Bildung, welche notwendige Voraussetzung für die Erreichung weiterführender klassischer Bildungsziele sowie den Erwerb einer vertieften Allgemeinbildung darstellen (vgl. Abschnitt 2.1.3). Ähnliche Argumente wurden bereits in den Anfängen der Diskussion um die Überprüfung von Bildungserträgen mittels standardisierter Leistungsmessung von den wichtigsten Vertretern der empirischen Bildungsforschung zu Beginn des neuen Jahrtausends vorgetragen (vgl. Baumert, 2001). Für den hier vorliegenden Kontext ist dabei die Idee zentral, dass eine vertiefte ästhetisch-expressive Begegnung und Gestaltung in den Bereichen Sprache und Literatur nur dann erfolgen kann, wenn die Grundlagen sprachlicher Kompetenz gesichert vorliegen (vgl. Köller, 2008, 2010). Hinreichende Basiskompetenzen stellen

somit die notwendige Voraussetzung für eine vertiefte und weiterführende Auseinandersetzung mit Sprache dar.

Ferner verweist Köller (2008, S. 50f.) darauf, dass die deutschen Bildungsstandards sehr stark in den Traditionen des jeweiligen Faches verankert sind und dass es daher den Fachdidaktiken obliege, das Verhältnis der jeweiligen fachspezifischen Standards zu einer solchen vertieften Allgemeinbildung herauszuarbeiten.

2.3 Der Kompetenzbegriff

Der für diese Dissertation zentrale Begriff *Kompetenz* (bzw. *Kompetenzkonstrukt*) besitzt zahlreiche verschiedene Ursprünge und – je nach Verwendungskontext und Wissenschaftsdisziplin – mannigfaltige Bedeutungsnuancen. Diese Vielfalt ist insofern problematisch, als der inflationäre und teils widersprüchliche Gebrauch den Begriff der „Kompetenz" zu einem unscharfen „Modebegriff der Sozial- und Erziehungswissenschaften" (Klieme & Hartig, 2007, S. 11) macht. Im aktuellen wissenschaftlichen Diskurs wird der Kompetenzbegriff oftmals unhinterfragt und ohne klar umgrenzten Bedeutungsgehalt verwendet, so dass dieser Gefahr läuft, als inhaltsleere Worthülse unbrauchbar zu werden. Auch kann die oft heftige Kritik der Allgemeinen Erziehungswissenschaft (vgl. bspw. Gruschka, 2006a, 2006b, 2007) an der empirischen Bildungsforschung sowie an den länderübergreifenden Bildungsstandards, die sich beide eng am Kompetenzbegriff orientieren, teilweise auf diese verschiedenen Ursprünge und divergierenden Bedeutungsnuancen zurückgeführt werden. Die eigentliche Intention der empirischen Erziehungswissenschaft beziehungsweise der empirischen Bildungsforschung war beziehungsweise ist es jedoch, mit dem Kompetenzbegriff den Grundstein für *präzise* Modellierungen und Messungen kognitiver Fähigkeiten und Fertigkeiten zu legen. Dies gründet auf der wissenschaftstheoretischen Sichtweise, dass die „klare und interpretationsresistente Formulierung wissenschaftlicher Fragen" (Hansel, 2007, S. 175f.) eine Grundvoraussetzung empirischer Hypothesenprüfung darstellt. „Diese wiederum ist nur auf der Grundlage scharfer Begriffe möglich. Solange man nur unklare Vorstellungen davon hat, was man sucht und demzufolge nur unklar formuliert, wohin man will, ist keine wissenschaftliche, demnach auch keine erziehungswissenschaftliche Erkenntnis möglich." (Hansel, 2007, S. 175 f.). Ferner verbindet sich mit dem Kompetenzbegriff und der Kompetenzorientierung die Hoffnung auf die Lösung aktueller pädagogischer Probleme. So bemerken auch Abraham und Kollegen (2007) aus deutschdidaktischer Perspektive: „Wo Defizite wahrgenommen werden, verspricht der Kompetenzbegriff die Lösung einer Reihe aktuell offenbar gewordener praktischer Probleme und eine neue Problemsicht" (Abraham et al., 2007, S. 6).

Um das in der vorliegenden Arbeit ausschlaggebende Verständnis des Kompetenzbegriffs abzuklären, soll zunächst auf die Ursprünge dieses Begriffs eingegangen werden. Vonken (2005, S. 18) unterscheidet sechs Konnotationen des Kompetenzbegriffs in verschiedenen fachwissenschaftlichen Disziplinen, welche neben der psychologischen und der linguistischen Konnotation, die nachfolgend ausführlich erläutert werden, auch die folgenden vier umfasst:

- Die soziologische Konnotation, die wesentlich von Max Weber geprägt wurde und Kompetenz im Rahmen früher Organisationstheorien als Zuständigkeit oder als die Verfügbarkeit von Machtmitteln versteht,
- die arbeitswissenschaftliche Konnotation, die Kompetenz als Kombination von „Dürfen" und „Können" sieht sowie
- eine betriebswirtschaftliche Konnotation, die dem Verständnis einer Verhalten generierenden Kompetenz entspricht.
- Ferner spricht Vonken (2005) gemäß seiner berufspädagogischen Orientierung von einer pädagogischen Konnotation, „die im Anschluss an die Schlüsselqualifikationsdebatte vor allem berufliche Handlungskompetenz thematisiert" (S. 18f.).

Die Vielfalt der Bedeutungsnuancen des Kompetenzbegriffs speist sich somit wesentlich aus seinen verschiedenen historischen Wurzeln, von denen für diese Dissertation insbesondere die linguistische sowie die psychologische Tradition relevant sind. Beide sollen nachfolgend näher betrachtet werden.

2.3.1 Sprachwissenschaftliches Begriffsverständnis

Einer der wichtigsten Ursprünge des Kompetenzbegriffs ist historisch eng mit der Erforschung sprachlicher Fähigkeiten verzahnt. So wurde der Begriff der *linguistic competence* bereits in den 1960er Jahren von dem vielzitierten Sprachwissenschaftler Noam Chomsky im Zusammenhang mit seiner Theorie der generativen Transformationsgrammatik geprägt. Er wählte diesen Begriff, um die allen Menschen gemeinsamen, kognitiven Wissensgrundlagen sprachlichen Handelns zu beschreiben. Chomsky (2006) äußert: „ […] we must isolate and study the system of linguistic competence that underlies behavior but that is not realized in any direct or simple way in behavior" (S. 4). Es gilt heute als Konsens, dass sich Chomsky mit dem Begriff competence ausschließlich auf (grammatisches) Wissen bezogen hat (vgl. Shohamy, 1996). Chomskys Arbeiten beinhalten die Sichtweise, dass das Sprachwissen von Menschen bzgl. ihrer

Muttersprache als prozedurales Wissen verstanden werden kann (vgl. Pollock & Cruz, 1999). Wie bei jeder Form prozeduralen Wissens gilt auch für Sprachwissen, dass Menschen die Regeln des Sprachgebrauchs internalisiert haben können, diese beim tatsächlichen Sprechen aber trotzdem verletzen. Daher ist die Intention der linguistischen Forschung, die internalisierten Regeln in einer Theorie *sprachlicher Kompetenz* zu beschreiben, nicht jedoch das tatsächliche, fehlerhafte Sprachhandeln, welches einer Theorie *sprachlicher Performanz* entsprechen würde. In Chomskys Sinne bezieht sich Kompetenz also auf die kognitiven Wissensgrundlagen, welche dem Menschen sprachliches Handeln – und damit Performanz – in immer neuen konkreten Situationen ermöglicht. Somit zielte Chomskys Begriffsverständnis aber nicht auf die unterschiedlichen „Kompetenzstände" von Individuen ab, wie sie im hier vorliegenden Kontext von Interesse sind, da solche Ungleichheiten seiner Ansicht nach lediglich die jeweilige situativ bedingte Performanz als aktuelle Manifestierung der zugrunde liegenden Kompetenz widerspiegeln. Auch Spolsky (1989, S. 138) betont, dass Chomsky den Kompetenzbegriff damit in einer deutlich anderen als der heute bildungswissenschaftlich gebräuchlichen Konnotation verwendete und man seine Lesart nicht mit der allgemein üblichen Bedeutung von Fähigkeit gleichsetzen dürfe.

In diesem Kontext ergibt sich eine grundlegende Schwierigkeit bei der Erfassung von Fähigkeiten, die in der Psychologie sowie in der Linguistik unter dem Stichwort Kompetenz-Performanz-Problem bekannt ist. In der psychologischen Kontrastierung von Kompetenz und Performanz kann Kompetenz als eine nicht direkt beobachtbare, zeitlich (relativ) stabile Fähigkeit verstanden werden, welche es einer Person gestattet, ein Problem zu lösen. Dementgegen ist Performanz ein direkt beobachtbares Verhalten in einer konkreten Situation, beispielsweise eine gegebene Antwort. In der Sprachwissenschaft bezieht sich die von Chomsky vorgenommene Unterscheidung von *competence* und *performance* auf die Abgrenzung der (grundlegenden) Sprachfähigkeit von der (aktuellen) Sprachverwendung. Unter Bezugnahme auf Bachman (1990) schreibt Shohamy (1996) zu diesem Thema: „ […] according to Bachman, there is a difference between competence and performance, where competence equals ability equals trait, while performance refers to the actual execution of tasks" (Shohamy, 1996, S. 148). In der quantitativ-empirischen Diagnostik resultiert aus dieser Unterscheidung das Problem, dass Personen mit vergleichbaren Kompetenzständen nur dann ähnliche (Test-)Leistungen erzielen, wenn sie auch ähnlich stark motiviert sind, ihre Leistungsdisposition in der Testsituation in ähnlicher Weise auszuschöpfen.

Nach Frederking (2010, S. 330) mündet die Differenz von Performanz und Kompetenz bei der Erforschung sprachlicher Phänomene in dem Ergebnis, dass Kompetenz den Forschungsgegenstand der Linguistik darstellt, während Performanz das Aufgabengebiet der Psycholinguistik oder Sprachpsychologie bestimmt. Dies sei nicht unproblematisch, da die

Psycholinguistik damit nur ein unselbständiger Appendix der Linguistik sei (vgl. Frederking, 2010, S. 330; Christmann & Groeben, 1997, S. 349).

Als Reaktion auf Chomskys linguistische Theorie wurde in den 1970er Jahren das Konzept der *communicative competence*, der kommunikativen Kompetenz eingeführt. Wegweisend waren hier beispielsweise die Arbeiten des Soziolinguisten Hymes (1972), der kommunikative Kompetenz nicht nur als grammatisches, sondern auch als sozio- und psycholinguistisches Wissen verstand. Neben der *Grammatikalität* der Form war für Hymes auch die *Angemessenheit* der Sprachverwendung relevant. Sein Konzept der kommunikativen Sprachkompetenz zeichnet sich durch die Ansicht aus, dass Regeln der Grammatik ohne die Regeln des Gebrauchs nutzlos wären. Hierbei definierte Hymes Kompetenz sehr allgemein und basal: „I should take *competence* as the most general term for the capabilities of a person" (Hymes, 1972, S. 282). Für das Konzept der kommunikativen Kompetenz ebenfalls bedeutsam war der Beitrag von Canale und Swain (1980), in dem eine auch heute noch gebräuchliche Definition des Begriffs der *communicative competence* und eine Beschreibung ihrer Bestandteile vorgestellt wurden. Hierbei liegt der Fokus nicht länger auf grammatikalisch wohlgeformten Sätzen, sondern vielmehr auf der in einer gegebenen Situation angemessenen Sprachverwendung, die zum Gelingen einer Kommunikation beiträgt.

Im deutschsprachigen Raum sind hinsichtlich des Konzepts der kommunikativen Kompetenz allerdings eher die Arbeiten von Habermas und seine Theorie des kommunikativen Handelns (1974/1989, 1981, 1982/1989) bekannt. In seinem Kompetenzmodell vereint Habermas identitätstheoretische und interaktionistische Aspekte und geht davon aus, dass es eine universelle, kulturübergreifende Kompetenz des sozialen Handelns gibt. Diese Kompetenz vereint in sich die drei Teilaspekte der kognitiven, der sprachlichen sowie der interaktiven Kompetenz (vgl. Habermas, 1974/1989).

Auch die Testung sprachlicher Kompetenzen lehnte sich während dieser Zeit an das Konzept der kommunikativen Kompetenz an und verfolgte das Ziel, sprachliche Leistungen in tatsächlich kommunikationsrelevanten Situationen zu erfassen. Später folgte die einflussreiche Arbeit von Bachman (1990), der sich an einer performanzbasierten Sichtweise sprachlicher Fähigkeiten orientiert. Die auf einer pragmatischen Sichtweise aufbauende Tradition der *performance tests* zeichnet sich zwar durch einen intuitiven Zugang, hohe Augenscheinvalidität und eine hieraus resultierende große Akzeptanz bei den Anwendern aus – die der sprachlichen Testleistung zugrunde liegenden kognitiven Prozesse stehen hier jedoch nicht länger im Fokus der Auseinandersetzung mit sprachlichen Kompetenzen. Andererseits äußert Bachman (1990): „A clear and explicit definition of language ability is essential to all language test development and use" (S. 3f.). Ferner beschäftigt sich Bachman intensiv mit Fragen der Mess- und Testtheorie. So

sind performanz- oder aufgabenbasierte Sprachtests zwar leicht zugänglich und erfüllen einerseits hohe testmethodische Qualitätsansprüche, entbehren aber andererseits in gewisser Weise den theoretischen Grundlagen, welche kognitiven Konstrukte hinter der beobachteten Performanz stehen.

2.3.2 Psychologisches Begriffsverständnis

In der Psychologie fand der Kompetenzbegriff erstmals in der Mitte des 20. Jahrhunderts durch den Motivationspsychologen Robert White (1959) Anwendung. Dieser definierte *competence* als „an organism's capacity to interact effectively with its environment" (White, 1959, S. 297). White ging davon aus, dass jedes Individuum intrinsisch motiviert sei, Kompetenz auszubilden, um dem Bedürfnis nach einer wirkungsvollen Interaktion mit seiner jeweiligen Umwelt gerecht werden zu können.

In den 70er Jahren des letzten Jahrhunderts kristallisierte sich die sozialpsychologische Sicht des Kompetenzbegriffs heraus, die sich in Abgrenzung zu einer dekontextualisierten Intelligenzdiagnostik für eine Diagnostik einsetzte, die sich an einer Bewährung in konkreten, im täglichen Leben relevanten Leistungssituationen orientiert (vgl. McClelland, 1973). Damit erhält der Kompetenzbegriff in der psychologischen Tradition eine neue Bedeutung, die der von Chomsky gewählten Definition entgegensteht. Klieme und Hartig (2007) bemerken in diesem Zusammenhang: „Als ‚Kompetenz' wird hier […] gerade das verstanden, was bei Chomsky und seinen Nachfolgern ‚Performanz' ist" (S. 16).

Häufig findet der Begriff der Kompetenz gemeinsam mit dem in der psychologischen Literatur verwendeten Begriffspaar *Fähigkeiten und Fertigkeiten* Anwendung. Bei dieser Unterscheidung bezeichnet der Begriff der Fähigkeit üblicherweise das angeborene oder erworbene (kognitive) Leistungsvermögen. Der Begriff der Fertigkeit wird im Allgemeinen in Zusammenhang mit erlernten beziehungsweise durch Übung ausgebildeten Verhaltenskomponenten gebraucht, wobei sich das Erlernte mitunter auf physische oder (fein-)motorische Aspekte wie beispielsweise Klavierspielen bezieht. Fähigkeit gilt als Voraussetzung für die Realisierung einer Fertigkeit. In aktuellen Begriffsbestimmungen umfassen Kompetenzen zumeist sowohl Fähigkeiten als auch Fertigkeiten (vgl. bspw. Weinert, 2001a).

Um der oben angesprochenen Bedeutungsdiffusion hinsichtlich des Kompetenzbegriffs entgegenzuwirken, widmen sich aktuelle Forschungsprojekte auf dem Gebiet der deutschsprachigen empirischen Bildungsforschung unter anderem der Schärfung dieser Begrifflichkeit (vgl. bspw. das DFG-Schwerpunktprogramm „Kompetenzmodelle zur Erfassung

individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen", Klieme & Leutner, 2006).

Hierbei steht die Frage im Vordergrund, wie der Kompetenzbegriff pointiert gefasst werden kann, ohne sich primär auf die von Weinert (2001a) vorgenommene Definition zurückzuziehen, die in ihrer messmethodischen Umsetzung problematisch ist. In seiner Definition beschreibt Weinert (2001a) Kompetenzen als „die bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen, sowie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten, die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können" (S. 27). Oftmals wird diese Definition verkürzt wiedergegeben, wobei allein auf die kognitiven Anteile des Kompetenzbegriffs verwiesen wird. Hierbei sind Kompetenzen in Kontrastierung zu allgemeinen, kontextunabhängigen kognitiven Fähigkeiten (Intelligenz) stets kontextspezifisch und zielen auf kognitive Anforderungen in spezifischen Bereichen (Koeppen et al., 2008). Aus dieser Gegenüberstellung resultiert auch, dass Kompetenzen – im Gegensatz zur allgemeinen kognitiven Leistungsfähigkeit, die durch Lernen kaum beeinflusst werden kann – durch (schulisches) Lernen und Erfahrungsbildung in den jeweiligen Bereichen aufgebaut werden können und sollen. Diese Unterscheidung von domänenspezifischen, erlernbaren Kompetenzen und allgemeiner kognitiver Leistungsfähigkeit (*gf*) ist von großer Bedeutung und wurde in den letzten Jahren teils kontrovers diskutiert. So meinen einzelne Wissenschaftler empirisch belegen zu können, dass internationale Schulleistungsstudien und Intelligenztest ein einziges latentes Fähigkeitskonstrukt messen, welches praktisch identisch mit allgemeiner Intelligenz ist (vgl. Rindermann, 2006, 2007). Wäre dies tatsächlich der Fall, wäre die Unterscheidung zwischen allgemeiner (*g*) beziehungsweise fluider Intelligenz (*gf*) und Schulleistung überflüssig. Diese These wurde auf nationaler sowie internationaler Ebene heftig und kontrovers diskutiert (vgl. Open Peer Commentary, 2007). Die Mehrheit der Experten innerhalb der deutschsprachigen empirischen Bildungsforschung lehnt diesen Standpunkt allerdings ab (vgl. bspw. Baumert, Brunner, Lütdke & Trautwein, 2007; Baumert, Lütdke, Trautwein & Brunner, 2009) und bietet alternative Erklärungen für den hohen, empirisch ermittelten Zusammenhang der Leistung in verschiedenen Kompetenzbereichen beziehungsweise Testdomänen (Lesen, Mathematik, Naturwissenschaften) und allgemeiner Intelligenz an.[2]

Das von Weinert (2001a, 2001b) vorgeschlagene Verständnis des Kompetenzbegriffs geht aber zunächst über kognitive Aspekte hinaus und umfasst zusätzlich motivationale, volitionale und soziale Bereitschaften und Fähigkeiten, also beispielsweise auch Einstellungen und

[2] Eine detaillierte Darstellung findet sich bspw. bei Baumert et al. (2007) sowie bei Baumert et al. (2009).

Erwartungen. Eine Trennung der kognitiven von den motivationalen Bestandteilen im Rahmen empirischer Untersuchungen wurde allerdings von Weinert selbst in die Diskussion eingebracht (vgl. Weinert 2001b). *Kompetenz* ist in diesem Verständnis *rein kognitiv* gefasst, die Verknüpfung mit motivationalen, volitionalen und sozialen Aspekten führt zum Begriff der *Handlungskompetenz*, die eine verantwortliche und erfolgreiche Nutzung von Kenntnissen und Fertigkeiten gestattet (Weinert, 2001a, S. 28).

Dem kognitionsorientierten Kompetenzkonzept entsprechen aktuelle Definitionen wie die von Koeppen und Kollegen (2008): „[…] we define competencies as context-specific cognitive dispositions that are acquired and needed to successfully cope with certain situations or tasks in specific domains" (Koeppen et al., 2008, S. 62). Zentral sind für den hier gebrauchten Kompetenzbegriff also zum einen der Kontextbezug und zum anderen die Erlernbarkeit (vgl. auch Klieme & Hartig, 2007, S. 18). Ferner stellen Kompetenzen eine „Verbindung von Wissen und Können in der Bewältigung von Handlungsanforderungen dar" (Klieme & Hartig, 2007, S. 19). Außerdem werden Kompetenzen regelmäßig als Verhaltens*dispositionen* konzipiert (vgl. Klieme & Hartig, 2007; Koeppen et al., 2008). Gegenwärtig wird in der einschlägigen empirischen Forschung zumeist eine so gefasste, kognitiv orientierte Definition des Kompetenzbegriffs zugrunde gelegt und in Testinstrumenten umgesetzt (vgl. Klieme & Leutner, 2006; Klieme & Hartig, 2007; Koeppen et al., 2008).

Für ein solches, auf kontextspezifische kognitive Leistungsdispositionen abzielendes Verständnis von Kompetenz ist die Tatsache relevant, dass aus beobachtbarem Verhalten nicht unmittelbar und fehlerfrei auf den Grad der Ausprägung einer zugrunde liegenden Kompetenz geschlossen werden kann. Kompetenzen werden somit als nicht direkt beobachtbare, latente Konstrukte verstanden, die über Indikatoren, beispielsweise Testaufgaben, operationalisiert und der Überprüfung zugänglich gemacht werden müssen. Dies wird im nachfolgenden Abschnitt 2.4 eingehender erläutert.

2.3.3 Kritische Diskussion des kognitionsorientierten Kompetenzbegriffs

Die Reduktion von Kompetenz auf prüfbare, kognitive Aspekte steht insbesondere im Rahmen der Erfassung sprachlicher Kompetenzen, aber auch unter Bezugnahme auf einen klassischen Bildungsbegriff in der Kritik (Gruschka, 2006a, 2006b, 2007; Merkens, 2006; Switalla, 2002). In unterschiedlichen Kontexten wurde beklagt, dass die alleinige Konzentration auf den kognitiven Kompetenzanteil ohne gleichzeitige Berücksichtigung von motivationalen, volitionalen und sozialen Bestandteilen zumeist ein nur unbefriedigendes Konstruktverständnis gestattet. So findet sich mitunter auch aus deutschdidaktischer Perspektive die Ansicht, dass die Testung sprachlicher

Kompetenzen im Rahmen von Large-Scale-Assessments insofern defizitär sei, als sie bestimmte Aspekte beispielsweise des Verstehens literarischer Texte nicht abbilden könne (vgl. bspw. Spinner, 2003, 2008). So schränke auch PISA das Konzept der Lesekompetenz auf kognitive Dimensionen ein und vernachlässige die ebenfalls zentralen emotional-motivationalen Aspekte. Gerade innerhalb solcher emotional-motivationaler Dimensionen könne man aber bestimmte Lern- und Entwicklungsprozesse generell nicht mit Leistungstests prüfen, insbesondere dann nicht, wenn diese auf geschlossene Aufgabenformate zurückgreifen (vgl. Groeben, 2005). Dass diese Sichtweise im Kontext sprachwissenschaftlicher Forschung eine lange Tradition besitzt, zeigt ein Zitat von Hymes, der bereits 1972 betonte: „In speaking of competence, it is especially important not to separate cognitive from affective and volitive factors, so far as the impact of theory on educational practice is concerned; but also with regard to research design and explanation […]" (Hymes, 1972, S. 283).

Diese Kritik hat aus meiner Sicht Berechtigung: Die testdiagnostische Messung sprachlicher Kompetenzen beschränkt sich oftmals auf kognitive Aspekte und drängt motivationale und affektive Facetten in den Hintergrund, obwohl diese fraglos zum Entstehen der erfassten Leistung beitragen. Diese gegenwärtige Praxis erklärt sich jedoch nicht aus der Ignoranz quantitativ-empirisch arbeitender Diagnostiker, sondern aus dem Umstand, dass es bislang nicht gelungen ist, im Rahmen des Large-Scale-Assessments Methoden zu etablieren, die sowohl dem vollen inhaltlichen Gehalt der angezielten Konstrukte als auch den psychometrischen Anforderungen an deren Messung gerecht werden können.

Leider wird aber auch das oben detailliert vorgestellte, gegenwärtige Verständnis des Kompetenzbegriffs in der Debatte um Bildungsstandards und Schulleistungstests mitunter nicht zur Kenntnis genommen. Wenn aus Sicht der Erziehungswissenschaft von Gruschka (2007) bemerkt wird: „Urteilt man begrifflich strenger, so kann man feststellen, dass es in dem realen Unternehmen *nicht mehr um Kompetenz* geht, sondern *schlicht um Performanz*, nämlich diejenigen, mehr oder weniger intelligente, verständliche und herausfordernde Aufgaben zu lösen" (S. 26) und Frederking (2010) als Vertreter der Deutschdidaktik dieser Sichtweise mit der Bemerkung zustimmt: „Mit Bezug auf die von Chomsky getroffene Unterscheidung zwischen Performanz und Kompetenz lässt sich nämlich berechtigt einwenden, dass die Bildungsstandards keine Kompetenz abbilden, sondern lediglich Performanz" (S. 334), muss beiden entgegnet werden, dass hier auf der Grundlage einer nicht zutreffenden Prämisse argumentiert wird: Das in der Psychologie und der empirischen Bildungsforschung vorherrschende Verständnis des Kompetenzbegriffs deckt sich *nicht* mit dem Begriffsverständnis von Chomsky. Um dies zu unterstreichen, setzen sich Vertreter der pädagogisch-psychologischen Diagnostik sowie der empirischen Bildungsforschung in zahlreichen Publikationen vertiefend mit den historischen

Wurzeln und dem gegenwärtigen Verständnis des Kompetenzbegriffs auseinander (vgl. bspw. Klieme & Hartig, 2007; Klieme, Hartig & Rauch, 2008; Koeppen et al., 2008) und bemühen sich, eine möglichst klare Definition des gegenwärtig gebrauchten Kompetenzbegriffs vorzunehmen. Auch wenn die von Chomsky geleisteten Arbeiten in diesem Zusammenhang stets gewürdigt werden, ist das aktuell vorherrschende Kompetenzverständnis, welches auch den Bildungsstandards zugrunde liegt, dennoch ein *anderes* als das von Chomsky verwendete. Oftmals wird dies auch entsprechend zur Kenntnis genommen. So bemerken beispielsweise aus deutschdidaktischer Sicht Abraham und Kollegen (2007): „Die Kompetenz fasste bei Chomsky die reine Anlage angeborener sprachlicher Fähigkeiten – und kein Lernergebnis. Das sieht die aktuelle Diskussion anders." (Abraham et al., 2007, S. 8). Mit Köller (2008) soll an dieser Stelle noch einmal betont werden, dass in einem psychologisch geprägten Verständnis Kompetenz nicht mit dem (im Test gezeigten Antwort-)Verhalten, also Performanz gleichgesetzt wird. Vielmehr bezeichnen Kompetenzen die von den Schülerinnen und Schülern erlernbaren beziehungsweise bei ihnen verfügbaren Fähigkeiten und Fertigkeiten, die erforderlich sind, bestimmte (Test-)Aufgaben lösen zu können (vgl. Köller, 2008, Porsch, 2010).

Ebenfalls kritisch diskutiert wird das Verhältnis der Begriffe *Kompetenz* und *Bildung* (vgl. Abschnitt 2.1.3). Da sich *Kompetenzen* im Sprachgebrauch der empirischen Bildungsforschung tendenziell und in der testdiagnostischen Umsetzung zumeist ausschließlich auf kognitive Aspekte beschränken, wird gegenüber dem Bildungsbegriff die Auslassung von affektiv-motivationalen Bestandteilen und der Verlust von Werthaltungen beklagt (vgl. bspw. Gruschka 2006a, 2006b, 2007). Da Bildung ein Prozess der Herausbildung von eigenverantwortlicher Mündigkeit und Selbstentfaltung sei, könne dieser nicht mit dem Kompetenzerwerb im Sinne des Erlernens und Aneignens von Fähigkeiten, Fertigkeiten und Wissensbeständen gleichgesetzt und dürfe nicht auf diesen reduziert werden. Dass eine gewisse inhaltliche Nähe zwischen dem Bildungsbegriff und der Kompetenzkonzeption besteht, wird von Vertretern der Allgemeinen Erziehungswissenschaft teilweise mit Nachdruck bestritten (vgl. bspw. Gruschka 2006a, 2006b, 2007).

Aus Perspektive der empirischen Pädagogik beziehungsweise der empirischen Bildungsforschung ist die Differenz zwischen Bildung und Kompetenz allerdings weniger augenscheinlich. Historisch greift erstmals Roth (1971, S. 180) Aspekte beider Begriffe auf und beschreibt „Mündigkeit als Kompetenz für verantwortliche Handlungsfähigkeit" als zentrales Erziehungsziel. Mündigkeit ist ein wesentlicher Aspekt des klassischen Bildungsbegriffs und wird von Roth als Ablösung der Fremdbestimmung durch möglichst weitgehende Selbstbestimmung verstanden (vgl. Roth, 1971, S. 180). Hierbei fasst Roth Kompetenzen als individuelle Handlungsdisposition auf, die im Verlauf von Bildungs- und Erziehungsprozessen erworben

werden und legt ein breites Begriffsverständnis zugrunde, welches nicht nur kognitive Aspekte berührt, sondern auch affektiv-motivationale Facetten einschließt. Die als Kompetenz interpretierte Mündigkeit bezieht Roth auf die drei Bereiche der Selbstkompetenz, der Sachkompetenz sowie der Sozialkompetenz (vgl. Roth, 1971, S. 180), welche sich als nachhaltig einflussreiche Unterscheidung herauskristallisierte (vgl. Oelkers & Reusser, 2008). Das pädagogische Ziel der Kompetenzvermittlung ist nach Roth „die Befähigung zu selbständigem und selbstverantwortlichem Handeln und damit zur Mündigkeit" (Klieme & Hartig, 2007, S. 21). Auf der Ebene der Begriffsfundierung finden also durchaus Aspekte der klassischen Bildungstheorie Eingang in die Kompetenzkonzeption, auf der Ebene der messtheoretischen Umsetzung der Konstrukte in Testinstrumente erfolgt allerdings notwendigerweise eine Konkretisierung – und damit eine Reduktion – der Kompetenzkonstrukte. Kompetenzen sind zwar keinesfalls identisch mit klassischen Bildungszielen, aber durchaus mit ihnen kompatibel und können einen Beitrag zu ihrer Realisierung leisten (vgl. Klieme & Hartig, 2007, S. 22).

2.3.4 Der Begriff der Sprachkompetenz

Wie in der oben erörterten Gegenüberstellung von Performanz und Kompetenz bereits angedeutet wurde, waren Bemühungen zur Definition sprachlicher Kompetenz lange Zeit von der Unterscheidung von *sprachlichem Wissen* und *sprachlichem Handeln* geprägt. Der heute prominente, allgemeine Kompetenzbegriff kann auf den Begriff der Sprachkompetenz übertragen werden und ist geeignet, eine Verbindung zwischen sprachlichem Wissen und seiner Anwendung im Sprachhandeln herzustellen und somit beide Aspekte zu integrieren. Hymes (1972) präzisiert im Kontext der kommunikativen Kompetenz entsprechend: „Competence is dependent upon both (tacit) *knowledge* and (ability for) *use*" (S. 282).

Eine einheitliche, allgemein akzeptierte und in diesem Sinne verbindliche Definition von Sprachkompetenz gibt es aber nicht (vgl. Jude, 2008). Vielmehr sind die existierenden Begriffsbestimmungen von den theoretischen Ansätzen und den jeweiligen Schwerpunktsetzungen der einzelnen Disziplinen gekennzeichnet, in denen sie zum Einsatz kommen (vgl. Abschnitt 2.5).

In der hier vorliegenden Arbeit steht allerdings weniger die Sprachkompetenz als übergreifende Schlüsselkompetenz (vgl. Weinert, 2001a) als vielmehr die Betrachtung spezifischer Aspekte beziehungsweise Bestandteile von Sprachkompetenz im Zentrum des Interesses. Diese Bestandteile werden nachfolgend als *sprachliche Kompetenzen* gefasst und entsprechen im Wesentlichen den in den Bildungsstandards für das Fach Deutsch im Primarbereich beschriebenen Kompetenzbereichen (KMK, 2004, 2005a, 2005b). Diese Vorgehensweise weicht

von der für das Fach Mathematik gewählten Konzeption ab, in der neben spezifischen Kompetenzen auch von einer allgemeinen mathematischen Kompetenz ausgegangen wird. Dieser Entscheidung liegt die Überlegung zugrunde, dass die schulischen – und auch außerschulischen – Lerngelegenheiten für mathematische Inhalte und Themen für die einzelnen Kompetenzbereiche weniger heterogen sind, als dies für die Kompetenzbereiche im Fach Deutsch angenommen werden muss. Die Unterscheidung einer umfassenden Kompetenz in der Verkehrssprache Deutsch von verschiedenen, differenzierbaren Kompetenzbereichen entspricht somit der Überlegung, dass Kompetenzen in der psychologischen Begriffstradition als „erlernbare und kontextspezifische Leistungsdispositionen" verstanden werden, „die sich funktional auf Situationen und Anforderungen in bestimmten Domänen beziehen" (Klieme & Hartig, 2007, S. 17), wobei die Breite der jeweils betrachteten Domäne erheblich variieren kann. Für eine als Schlüsselkompetenz verstandene umfassende Sprachkompetenz – die sich möglicherweise sogar auf verkehrs- *und* fremdsprachliche Facetten erstreckt – wäre die anzunehmende Domänenbreite sowie Anzahl und Heterogenität der involvierten Kompetenzaspekte sehr groß und eine erschöpfende Operationalisierung entsprechend schwierig umzusetzen. Ferner ist anzunehmen, dass sich die Leistungsstände in den relevanten Kompetenzbereichen in der Verkehrssprache Deutsch (Zuhören, Lesen, Sprechen, Schreiben) im Primarbereich intraindividuell aufgrund der jeweils verschieden weit fortgeschrittenen kindlichen Sprachentwicklung (vgl. Abschnitt 2.6) noch sehr stark unterscheiden. Für mündliche Kompetenzen standen zu diesem Zeitpunkt bereits mehrjährige, auch außerschulische Lerngelegenheiten zur Verfügung, während die schriftsprachlichen Kompetenzen erst am Beginn ihrer primär schulisch geförderten Ausbildung stehen. Die Konzeption einer umfassenden Sprachkompetenz müsste für Grundschulkinder also die deutlich divergenten Entwicklungsstände für verschiedene relevante Kompetenzbestandteile berücksichtigen, welche in der Sekundarstufe nicht mehr in derselben Weise vorliegen. Folgt man dieser Überlegung, ist eine für die gesamte Schulzeit anwendbare, jahrgangsübergreifende Konzeption von Sprachkompetenz kaum umsetzbar. Deshalb wird in dieser Dissertation keine Definition und Operationalisierung einer umfassenden Sprachkompetenz angestrebt, sondern das Augenmerk auf jeweils spezifische, den Kompetenzbereichen der länderübergreifenden Bildungsstandards entsprechenden sprachlichen Kompetenzen gelegt.

Den Beziehungen dieser sprachlichen Kompetenzen und ihrem Verhältnis zu einer möglichen umfassenden Sprachkompetenz wird aus theoretischer und empirischer Sicht in Abschnitt 3.1 nachgegangen. Eine vertiefende Definition der einzelnen sprachlichen Kompetenzen erfolgt in den jeweiligen Abschnitten beziehungsweise in den Einzelbeiträgen dieser Arbeit.

2.4 Die Operationalisierung von latenten Kompetenzkonstrukten und messmethodische Grundlagen

2.4.1 Überlegungen zur Konstruktvalidierung

Kompetenzen werden als Dispositionen verstanden, die gezeigtem Verhalten und somit auch Leistungen zugrunde liegen. Leistungen sind die Ergebnisse von Handlungen, die nach einem Gütemaßstab bewertbar sind. Allerdings muss bei der Bewertung und Interpretation von Leistungsergebnissen berücksichtigt werden, dass die in einer bestimmten (Test-)Aufgabe erzielte Leistung sowohl von der Kompetenzausprägung als auch von der aktuellen Anstrengungsbereitschaft abhängt. Leistungsunterschiede zwischen Personen können also nur unter Annahme gleicher Anstrengungsbereitschaft als Unterschiede in der Kompetenzausprägung verstanden werden.

Möchte man sprachliche Kompetenzkonstrukte einer empirischen Messung zugänglich machen, ist in einem ersten Schritt eine möglichst präzise Definition der fraglichen Kompetenz zu leisten, die beispielsweise in Form eines Kompetenzmodells Ausdruck finden kann. Anschließend muss spezifiziert werden, in welchen Situationen sich intra- wie interindividuelle Unterschiede in der Kompetenzausprägung in welchen Verhaltensäußerungen niederschlagen. Erst durch eine solche Spezifikation kompetenzrelevanter Situationen können Testinhalte bestimmt werden, die eine Differenzierung von kompetentem und nicht kompetentem Antwortverhalten gestatten. Konstruktdefinition und Spezifikation von kompetenzrelevanten (Test-)Situationen sind somit die Voraussetzung für die Operationalisierung eines interessierenden Kompetenzkonstrukts in einem Messverfahren (vgl. Klieme & Hartig, 2007). Möchte man nun die Ergebnisse eines sprachlichen Leistungstests als Grundlage verwenden, um Rückschlüsse über die sprachliche Kompetenz eines Testteilnehmers zu ziehen, muss ferner abgeklärt werden, in welcher Beziehung die Leistung im Sprachtest zu anderem, nicht testbezogenem Sprachgebrauch steht. Es muss also ein theoretischer Rahmen definiert sein, der es gestattet, die sprachbezogene Testleistung als einen spezifischen Fall der fraglichen sprachlichen Kompetenzäußerung zu betrachten.

Bachman und Palmer (1996) postulieren in diesem Zusammenhang, dass es notwendig ist, eine klare Entsprechung von Eigenschaften der zu beurteilenden sprachlichen Verwendungssituation und Eigenschaften des Sprachtests herzustellen, um aus der Testperformanz verlässliche Schlüsse hinsichtlich der zugrunde liegenden sprachlichen Kompetenz ableiten zu können.

Zusammenfassend schreiben Koeppen und Kollegen (2008): „In short, the measurement of competencies should be based on a solid theoretical and psychometric basis that allows the measurement result (e.g., quantity and quality of solved tasks) to be interpreted with reference to an underlying theoretical model of competencies" (Koeppen et al., 2008, S. 62).

Diese hier dargelegten Überlegungen berühren die Frage nach der Validität der Messung beziehungsweise der Interpretation eines Testergebnisses[3]. Um hier eine Antwort zu finden, ist es erforderlich zu belegen, dass ein interessierendes Testergebnis tatsächlich den Bereich beziehungsweise die Bereiche sprachlicher Fähigkeit widerspiegelt, die der jeweiligen Messintention entsprechen – und zwar *nur* oder zumindest *primär* dieser. Entsprechende Belege lassen sich allerdings nur dann finden, wenn das angezielte sprachliche Konstrukt eingangs klar genug definiert worden ist. Der Begriff der Konstruktvalidität bezieht sich somit auf das Ausmaß, in dem ein erhaltener Testwert als Indikator für die sprachliche Fähigkeit beziehungsweise das Kompetenzkonstrukt, das gemessen werden soll, interpretiert werden kann. Im Prozess der *Konstruktvalidierung* wird entsprechend nach Belegen dafür gesucht, dass die erfolgte Interpretation eines Testwerts – als Indikator für die Ausprägung einer bestimmten sprachlichen Kompetenz – gerechtfertigt ist.

Wegweisende Arbeiten zur Klärung der Frage, wie Konstrukte in Abgrenzung zu Aufgaben beschrieben werden können und wie sich Konstruktvalidität bestimmen lässt, stammen von Samuel Messick (1989a, 1994). Seiner Auffassung nach bezieht sich der Begriff des Konstrukts auf theoretische Vorstellungen zu Kompetenzen, Wissensbeständen und Fertigkeiten, die einer (Test-)Performanz zugrunde liegen. Aufgaben beziehen sich direkt auf das Antwortverhalten, also die (Test-)Performanz. Auch wenn versucht wird, aus Beobachtungen von Performanz auf Kompetenzausprägungen zurückzuschließen, so sind diese Schlussfolgerungen selten eindeutig und mitunter fehlerhaft, insbesondere dann, wenn aus einer schlechten Testperformanz auf eine geringe Kompetenzausprägung zurückgeschlossen wird. Von ausschlaggebender Bedeutung ist es, im Zuge der Konstruktvalidierung empirische Befunde so in die theoretischen Annahmen zu integrieren, dass ein konsistentes (Antwort-)Verhalten durch ein spezifisches Konstrukt und nicht durch plausible Alternativinterpretationen erklärbar ist. Insbesondere im Bildungskontext interessiert selten eine spezifische Testleistung, vielmehr richtet sich der Blick auf eine Kompetenz, welche nicht nur für eine einzelne Testleistung, sondern auch für eine Reihe anderer Testleistungen in ähnlichen Aufgaben als ursächlich gelten kann.

[3] In der aktuellen Diskussion des Validitätsbegriffs spricht man nicht länger davon, Aussagen bzgl. der Validität eines Tests anzustreben. Vielmehr steht die Validität der Interpretation von Testergebnissen und deren Nutzung im Fokus der Aufmerksamkeit (vgl. Messick, 1994). Hinsichtlich der Überprüfung der Testeigenschaften verwendet man den Begriff Validierung.

Die Charakteristik des Konstrukts sollte sowohl die Entwicklung und Auswahl der Aufgaben als auch die Kodieranweisungen und Auswertungskategorien bestimmen. Validität definiert Messick (1994) wie folgt: „Validity is an overall evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment" (S. 6). Messick betont, dass Validität keine Eigenschaft von Testverfahren als solchen ist, sondern sich auf die Bedeutung von Testergebnissen bezieht.

Die Fokussierung auf die kognitiven Grundlagen eines theoretisch durchdachten Konstrukts erhöht gleichzeitig die Aufmerksamkeit für konstruktirrelevante Einflüsse. Nach Messick (1994, S. 9) sind die zwei größten Bedrohungen der Konstruktvalidität die Phänomene der *construct underrepresentation* sowie der *construct-irrelevant variance*.

Das erste Phänomen tritt dann auf, wenn die Messung zu begrenzt angelegt ist und wichtige Facetten des Konstrukts ausgespart werden. Dieses Problem wird im hier vorliegenden Kontext der Bildungsstandards für das Fach Deutsch beispielsweise bei der Operationalisierung der Lesekompetenz offenbar, welche sich in der angloamerikanischen *reading literacy*-Tradition im Wesentlichen auf das Leseverstehen und den handlungskompetenten Umgang mit Textinformationen beschränkt. Aus Sicht der Literaturdidaktik werden dabei wesentliche Aspekte literarischer (Rezeptions-)Kompetenz nur unzureichend erfasst, was als Aussparung essentieller Anteile eines umfassenden Lesekompetenzkonstrukts gedeutet werden kann. Tatsächlich ist es besonders schwierig, literarische Kompetenz so in Aufgaben zu überführen, dass einerseits theoretischen Ansprüchen und andererseits messtheoretischen Notwendigkeiten entsprochen werden kann.

Das zweite von Messick (1994) thematisierte Phänomen der *construct-irrelevant variance* beschreibt das Problem, dass Operationalisierungen mitunter so erfolgen, dass das Antwortverhalten nicht allein durch das interessierende Konstrukt verursacht wird, sondern weitere Aspekte in die Messung einfließen. Hierbei kann es sich zum einen um messmethodische Artefakte wie beispielsweise das Rateverhalten bei Richtig-Falsch-Aufgaben handeln. Zum anderen können aber auch distinkte, inhaltlich sinnvolle, mit der aktuellen Messung allerdings nicht angestrebte Konstrukte das Antwortverhalten beeinflussen. Ein Beispiel hierfür ist die Beeinflussung der schriftlichen Leistungsüberprüfung von rezeptiven Kompetenzen, also dem Zuhören oder Lesen, durch die produktive Kompetenz des Schreibens. Zabka (2010) fasst dieses Problem wie folgt: „Am Schreibresultat ist häufig nicht eindeutig erkennbar, in welcher Mischung sich Kompetenzen des Textverstehens und der Textproduktion zeigen. […] Niemals werden Verstehenskompetenzen gültig getestet, stets mischen sich Schreibkompetenzen als Störvariablen ein" (S. 60).

2.4.2 Die Spezifikation von Messmodellen

Die in der vorliegenden Arbeit betrachteten sprachlichen Kompetenzkonstrukte sind insofern theoretischer Natur, als eine direkte Beobachtung ihrer Ausprägungen nicht möglich ist, weshalb auch von latenten Konstrukten oder latenten Variablen gesprochen wird. Um Aussagen bzgl. der interessierenden latenten Konstrukte machen zu können, ist ihre Operationalisierung mittels eines Messmodells erforderlich. Dieser Schritt der Operationalisierung beinhaltet, dass für die nicht messbaren Konstrukte messbare Indikatoren identifiziert beziehungsweise hergestellt werden, die einer direkten Beobachtung zugänglich sind. Im psychometrischen Sprachgebrauch werden in diesem Zusammenhang die Begriffe *latent* vs. *observed variables* verwendet. Diese können wie folgt beschrieben werden: „Observed variables are epistemically accessible to the researcher, whereas latent variables are not epistemically accessible" (Borsboom, 2008, S. 27). Eine so formulierte Unterscheidung ist aus Sicht der *Latent Variable Theory* allerdings nicht unproblematisch, da *beobachtete Variablen* in den seltensten Fällen tatsächlich aus einer direkten *Beobachtung* resultieren. Vielmehr werde sowohl für latente als auch für beobachtete Variablen aus einer gegebenen Datenmatrix ein Rückschluss auf die Ausprägung des zugrunde liegenden Merkmals gezogen. Entscheidend sei daher die Frage, ob ein Rückschluss von der Ausprägung der Daten auf die Ausprägung des Merkmals mit Unsicherheit behaftet ist oder nicht:

> When we treat a variable as *observed*, we mean nothing more than that we assume that the location of a person on that variable can be inferred with certainty from the data. When we treat a variable as *latent*, we mean that the inference in question cannot be made with certainty. (Borsboom, 2008, S. 30)

Das Ausmaß der Unsicherheit wiederum ist eine Funktion der jeweils vorliegenden Daten. Ob eine Variable nun als latent oder als beobachtet gelten muss, kann daher nicht als inhärente Eigenschaft dieser Variablen verstanden werden.

Von Seiten psychometrischer Experten wird mitunter die Kritik geäußert, dass die Operationalisierung von Konstrukten, also die Überführung eines latenten, nicht beobachtbaren Konstrukts in beobachtbare Indikatoren nicht hinreichend durchdacht wird (vgl. Borsboom, 2006). Viel zu oft werde psychologische Forschung und insbesondere die Analyse von Daten sowie die Interpretation der Analyseergebnisse so betrieben, als sei der beobachtete Testwert mit dem interessierenden Konstrukt identisch. Es fehle das Verständnis für die Tatsache, dass Eigenschaften eines ermittelten Testwerts nur unter sehr strengen Annahmen – die in der Praxis oftmals nicht erfüllt sind – direkt als Eigenschaften des Konstrukts aufgefasst werden könnten. Somit sei von entscheidender Bedeutung, Folgendes zu berücksichtigen: „[…] the philosophical idea that theoretical attributes are, in fact, distinct from a set of observations, i.e., that one rejects

the operationalist thesis that theoretical attributes are synonymous with the way they are measured" (Borsboom, 2006, S. 428).

Um sprachliche Kompetenzkonstrukte und ihre strukturellen Beziehungen einer empirischen Untersuchung zugänglich machen zu können, muss die individuelle Ausprägung des interessierenden Kompetenzkonstrukts möglichst genau bestimmt werden. Eine solche *Messung* der individuellen Kompetenzausprägung fasst möglichst viele Beobachtungen der Kompetenzäußerung zusammen. Entscheidend für die Qualität der Messung ist, dass einerseits mehrere, variierende Anlässe der Kompetenzäußerung einbezogen werden, diese verschiedenen Anlässe aber tatsächlich alle eine möglichst exklusive Äußerung der fraglichen sprachlichen Kompetenz provozieren. Im Rahmen großer Schulleistungsstudien beschränken sich Beobachtungsanlässe leider oftmals auf den Einsatz standardisierter Testinstrumente, da andere Formen der Verhaltensbeobachtung in großen Stichproben einerseits nicht mit vertretbarem Aufwand umsetzbar sind und andererseits häufig den Mindestanforderungen an die Messqualität nicht genügen können. Somit beschränkt sich die Realisierung der Forderung nach möglichst vielen, verschiedenen „Beobachtungen" also auf die Umsetzung innerhalb testkonstruktiver Möglichkeiten. Im Falle der Messung der Lesekompetenz genügt es für die Abschätzung einer individuellen Fähigkeitsausprägung also nicht, dem betreffenden Testteilnehmer eine einzige Verständnisfrage zu nur einem Stimulustext vorzulegen. Zielführender ist es, mehrere Fragen zu mehreren verschiedenen Texten zu stellen[4] und den Einfluss anderer (sprachlicher) Kompetenzen möglichst gering zu halten, also beispielsweise keine ausführlichen schriftlichen Antworten zu verlangen, da sich hierbei neben der Lesekompetenz auch die Schreibkompetenz in beträchtlichem Umfang niederschlagen würde (s. o.).

Hinterfragt man den Zusammenhang zwischen einem latenten Konstrukt und seinen Indikatoren, sind grundsätzlich zwei Arten der *Modellspezifikation* denkbar, die beide in der quantitativ-empirischen Forschung Verwendung finden, allerdings grundsätzlich verschiedene Implikationen bergen. Man spricht in diesem Zusammenhang von *formativ vs. reflektiv* spezifizierten Messmodellen (vgl. bspw. Bollen & Lennox, 1991). Eberl (2004) fasst unter den Begriff der Spezifikation in diesem Zusammenhang „[s]owohl die zu Grunde liegende ‚reale' Kausalbeziehung der Konstrukte und ihrer Indikatoren als auch die vom Forscher im Rahmen eines Messmodells hypothetisierte Kausalbeziehung […]" (S. 2).

[4] Allerdings wird auch das Ziel einer möglichst vielfältigen Operationalisierung und das Vorlegen möglichst vieler Einzelaufgaben je interessierendem Kompetenzbereich durch testadministrative Beschränkungen – bspw. die limitierte Testzeit – determiniert.

Bei der Spezifikation eines reflektiven Messmodells (vgl. Abbildung 1) gilt die Ausprägung der latenten Variablen als ursächlich für die Ausprägung der beobachtbaren Variablen. Eine Veränderung der Ausprägung der latenten Variablen führt unmittelbar zu einer Veränderung in der Ausprägung der beobachtbaren Variablen und zwar – unter Vernachlässigung der eventuell unterschiedlichen Reliabilitäten der Indikatoren – bei allen zu einer identischen Veränderung.

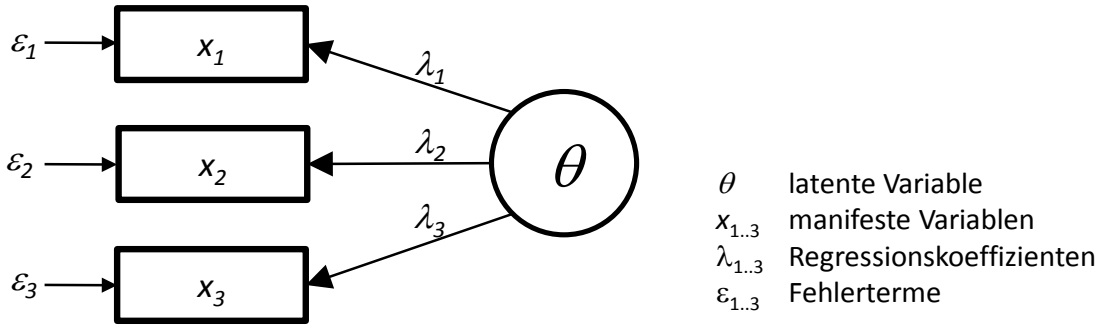

Abbildung 1: Reflektives Messmodel

Bei einem reflektiven Messmodell wird davon ausgegangen, dass für jedes latente Konstrukt durch seine Definition auch bestimmt werden kann, welche beobachtbaren Variablen konzeptionell durch diese Definition einbezogen werden und welche Indikatoren somit potentiell für die Messung des latenten Konstrukts geeignet sind. Somit besitzt jedes latente Konstrukt ein beschreibbares Indikatorenuniversum, aus dem ein unendlicher Itempool generiert werden könnte. Theoretisch sollten alle Items dieses Pools in derselben Weise geeignet sein, das Konstrukt zu messen. Ferner sollten sie möglichst hoch miteinander korreliert sein, da sie nur dann alle demselben Konstrukt entsprechen. Setzt man voraus, dass alle Indikatoren gleich reliabel messen, wären diese zudem beliebig austauschbar. Die hier dargestellte Idee eines reflektiven Messmodells entspricht dem Konzept des *domain-sampling* (Nunnally & Bernstein, 1994).

Dem entgegen ist die postulierte Beziehungsrichtung eines formativen Messmodells umgekehrt (vgl. Abbildung 2). Im Sinne eines strikten Operationalismus definieren hier die Indikatoren das latente Konstrukt.

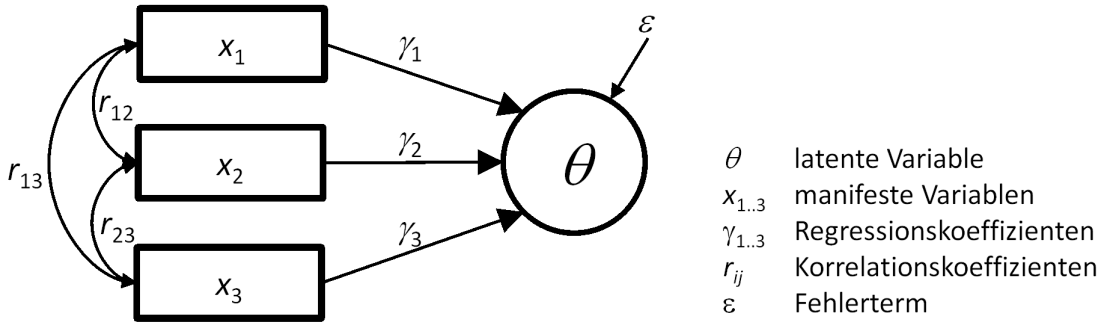

Abbildung 2: Formatives Messmodell

Die Indikatoren eines formativen Messmodells können, müssen aber untereinander nicht korreliert sein. Durch die Veränderung der Ausprägung eines Indikators ändert sich zwar die Ausprägung des latenten Konstrukts, die anderen Indikatoren können davon aber – je nach Stärke der korrelativen Beziehungen der Indikatoren untereinander – völlig unberührt bleiben. Übliche Verfahren zur Überprüfung der Messgüte basieren aber auf der Annahme, dass die Indikatoren untereinander korrelative Beziehungen aufweisen und können daher nicht auf formative Messmodelle angewendet werden.

Ein häufig zitiertes Beispiel für ein formatives Messmodell ist der sozioökonomische Status (SES), der sich durch die Einbeziehung der Indikatoren Bildung, Einkommen und Berufsprestige definiert und seinen inhaltlichen Gehalt verändern würde, wenn einer der Indikatoren weggelassen oder ersetzt werden würde. Indikatoren in einem formativen Messmodell sind somit nicht austauschbar.

Im Forschungsprozess wird nur selten bewusst reflektiert und dann aufgrund theoretischer Erwägungen begründet entschieden, ob ein formatives oder ein reflektives Messmodell angezeigt ist. Borsboom (2006) beklagt diesen Umstand mit Nachdruck:

> No attention is given to the crucial question how psychological attributes are structured, how they are related to observed scores, how one can utilize substantive theory in test construction, or how one can test one's ideas through the construction and evaluation of measurement models. (Borsboom, 2006, S. 436)

Auch wenn Borsbooms Feststellung in dieser Allgemeinheit sicher nicht zutrifft, wäre es in jedem Fall wünschenswert, *vor* Beginn der Testentwicklung ein latentes Konstrukt dahingehend zu hinterfragen, ob es treffender in einem formativen oder in einem reflektiven Messmodell spezifiziert werden sollte. Entscheidend ist in diesem Zusammenhang, dass die Entscheidung für

eine der beiden Optionen bewusst und überlegt erfolgt. Klagen, dass die eine Variante der Modellspezifikation gegenüber der anderen unreflektiert bevorzugt wird, gibt es in beide Richtungen (vgl. bspw. Bollen, 1989; Borsboom, 2006).

Für die hier vorliegende Arbeit lässt sich nun die Frage stellen, wie die Modellierung der Beziehung zwischen Konstrukt und Indikatoren im Kontext der Operationalisierung von sprachlichen Kompetenzkonstrukten erfolgen sollte. In der empirischen Bildungsforschung herrschen reflektive Modellspezifizierungen vor, wobei die latente Variable für gewöhnlich als kontinuierlich betrachtet und für die beobachtete Variable eine kategoriale (bspw. dichotome) Ausprägung angenommen wird. Dies führt innerhalb der Modellgruppe latenter Variablen zu Modellen der probabilistischen Testtheorie beziehungsweise der *Item Response Theory* (IRT) (für eine ausführliche Darstellung vgl. bspw. Embretson & Reise, 2000; Rost, 2004).

Modelle der probabilistischen Testtheorie beschreiben den Zusammenhang zwischen dem Lösungsverhalten bei der Aufgabenbearbeitung und dem, diesem Verhalten zugrunde liegenden, latenten Kompetenzkonstrukt als eine Wahrscheinlichkeitsfunktion. Ein wesentlicher Vorzug von Modellen aus der Klasse der IRT besteht darin, dass Personenfähigkeiten und Aufgabenschwierigkeiten auf derselben Skala abgetragen werden und somit die spezifischen situativen Anforderungen von Aufgaben, die einem bestimmten Niveau der Kompetenzausprägung entsprechen, unmittelbar zugeordnet werden können. Dies ist für die Entwicklung von Kompetenzstufenmodellen und die Beschreibung der darin unterschiedenen Niveaustufen von ausschlaggebender Bedeutung.

2.4.3 Konsequenzen der Modellspezifikation am Beispiel der Lesekompetenz

Die oben dargestellten allgemeinen Überlegungen zur Spezifikation von Messmodellen sollen an dieser Stelle für das Konstrukt der Lesekompetenz vertiefend betrachtet werden.

Das Konstrukt der Lesekompetenz wurde aus psychologischer Perspektive in den letzten Jahren beziehungsweise Jahrzehnten äußerst gründlich untersucht und beschrieben (vgl. bspw. Christmann, 2010; Groeben & Hurrelmann, 2002) und auch die Konstruktdefinition in großen nationalen und internationalen Schulleistungsstudien erfolgte sehr detailliert und bemühte sich um eine vertiefte theoretische Fundierung (vgl. Kintsch, 1998; Kirsch, Jungeblut & Mosenthal, 1998). Somit kann festgestellt werden, dass die theoretische Durchdringung und testdiagnostische Operationalisierung der Lesekompetenz bereits weit vorangeschritten ist und sich in dieser Hinsicht von anderen sprachlichen Kompetenzkonstrukten abhebt.

Das Ergebnis bisheriger Forschungsbemühungen und Testentwicklungen war stets die Spezifizierung reflektiver Messmodelle, denen – wie oben erläutert – die Annahme zugrunde

liegt, dass die Ausprägung der beobachteten Variablen (in diesem Fall die Itemantworten) kausal durch die Ausprägung der latenten Variablen (in diesem Fall die Lesefähigkeit) verursacht wird. Eine Veränderung, beispielsweise ein Anstieg, der schülerseitigen Lesefähigkeit müsste sich demnach auf *alle* Items, die der Messung der Lesefähigkeit dienen, *in gleicher Weise* auswirken. Je nach Konzeption der jeweiligen Studie werden nun in die Messung der Lesefähigkeit Items zu verschiedenen Arten von Lesetexten einbezogen. Dies betrifft mindestens die Unterscheidung von literarischen und Sach- beziehungsweise informierenden Texten (vgl. für IGLU 2006: Bos et al., 2007) und erstreckt sich vereinzelt auf die zusätzliche Unterscheidung von kontinuierlichen und diskontinuierlichen Lesetexten (vgl. für PISA 2000: Artelt, Stanat, Schneider & Schiefele, 2001; vgl. für den Ländervergleich Sprachen 2009: Böhme, Neumann & Schipolowski, 2010). Im Rahmen der Operationalisierung der Bildungsstandards im Fach Deutsch für den Primarbereich wurden im Gegensatz zur IGLU-Studie alle drei Textarten (literarische sowie kontinuierliche und diskontinuierliche Sachtexte) einbezogen (vgl. Pietsch, Böhme, Robitzsch & Stubbe, 2009). Geht man nun davon aus, dass als normative Setzung das Verstehen aller drei Textarten Bestandteil der Lesekompetenz ist, obwohl diese drei Komponenten hinsichtlich der ihnen zugrunde liegenden kognitiven Prozesse nicht vollständig identisch sind, kann die Verschiedenartigkeit der eingesetzten Lesestimuli in einer reflektiven Spezifikation des Messmodells problematisch sein. Dies lässt sich aus zwei Perspektiven betrachten:

Zum einen sind die Anteile an Testitems, die sich auf Stimulustexte einer bestimmten Textart beziehen, mitunter nicht gleich verteilt. Während für die Messung der Lesekompetenz im Rahmen der Operationalisierung der Bildungsstandards im Fach Deutsch für den Primarbereich eine annähernde Gleichverteilung der Anzahl der Stimulustexte über die Textarten hinweg angestrebt und erreicht wurde (vgl. Böhme & Robitzsch, 2009a, S. 259), sind beispielsweise im Rahmen der PISA-Studie 2000 die Itemzahlen für die verschiedenen Textarten deutlich divergent. So berichten Artelt und Schlagmüller (2004), dass in ihre Analysen zur Dimensionalität der Lesekompetenz 70 Items zu kontinuierlichen Sachtexten und 42 Items zu nichtkontinuierlichen Sachtexten, allerdings nur 17 Items zu (kontinuierlichen) literarischen Texten einbezogen wurden.[5] Im Lesetest der PISA-Studie des Jahres 2000 lag also keine Gleichgewichtung von Items zu den verschiedenen Textarten vor. Ebenso scheint es denkbar, dass beispielsweise im Anschluss an die von Böhme und Robitzsch (2009a) berichteten Befunde zum differentiellen Funktionieren von Leseitems zu komplexen Tabellen zwischen den Jahrgangsstufen drei und vier solche Stimuli in einem Pool von Leseitems für den Primarbereich

[5] Der Hintergrund dieser Ungleichgewichtung ist der Umstand, dass es deutlich schwerer ist, literarische Texte zu finden, die der Überprüfung interkultureller Brauchbarkeit sowie den verschiedenen Übersetzungsschritten standhalten können und somit für den internationalen Einsatz geeignet erscheinen (vgl. Artelt & Schlagmüller, 2004).

künftig in geringerem Umfang berücksichtigt werden. Ohne entsprechenden Ersatz wären Items zu diskontinuierlichen Sachtexten dann allerdings unterrepräsentiert. Erinnern wir uns an dieser Stelle kurz an die oben eingeführte Prämisse, dass die Lesekompetenz für die verschiedenen Textarten (kontinuierliche literarische Texte, kontinuierliche Sachtexte sowie diskontinuierliche Sachtexte) keine vollständig homogene und somit eindimensionale Struktur[6] aufweist. Dann könnte – nach einer ersten empirischen Erprobung des Itempools – eine Itemselektion nach den in der empirischen Bildungsforschung mittlerweile üblichen Kriterien (siehe hierzu bspw. Winkelmann & Böhme, 2009, S. 37), zur Aussonderung derjenigen Items führen, die im Pool unterrepräsentiert sind. Dies bedeutet nun aber keinesfalls, dass beispielsweise die Items zur Messung der Lesekompetenz bei literarischen Texten keine Überprüfung der Lesekompetenz gestatten würden. So schreibt auch Robitzsch (2009): „Itemselektionsprozesse, die auf Eindimensionalität von θ abzielen, behalten eher homogene Items der Teilbereiche bei und decken demzufolge nicht die gesamte Breite des zu repräsentierenden Itempools zulasten der Validität ab“ (S. 51).

Zum anderen zeigen die soeben erwähnten Analysen zum differentiellen Funktionieren von Leseitems zwischen den Jahrgangsstufen drei und vier, dass nicht davon ausgegangen werden kann, dass die Entwicklung der Lesekompetenz für alle betrachteten Stimuli *homogen*, also in genau gleicher Weise erfolgt. Dies ist jedoch eine grundlegende Annahme reflektiver Messmodelle.

Eine mögliche Lösung dieses Problems könnte die von Robitzsch (2009) für die Kompetenzmessung im Rahmen von Bildungsstandards vorgeschlagene Spezifizierung eines *Second Order Formative Models* darstellen (vgl. Robitzsch, 2009, S. 50f.). Ein solches Modell kann auf der ersten Ebene als reflektives Messmodell verstanden werden, bei dem Testitems jeweils verschiedene (latente) Konstruktbestandteile messen, so beispielsweise die Lesekompetenz bei literarischen sowie kontinuierlichen und diskontinuierlichen Sachtexten. Auf der übergeordneten zweiten Ebene dienen diese Teilkompetenzen dann formativ zur Definition eines allgemeinen Lesekompetenzkonstrukts. Dieses wäre dann *normativ* definiert, da es der a priori erfolgten Setzung entspricht, sowohl das Verständnis kontinuierlicher literarischer und Sachtexte als auch das Verständnis diskontinuierlicher Sachtexte abzubilden. Auf diese Weise ließe sich das Problem handhaben, dass bei einem rein reflektiven Messmodell nach „traditionellen“ Kriterien der Itemselektion mitunter solche Items als inhomogen abgelehnt werden würden, die im Itempool unterrepräsentiert sind. Ferner könnten die – je nach Textart – möglicherweise unterschiedlichen Zuwächse in der Lesekompetenz mit den Modellannahmen in Einklang gebracht werden.

[6] Zur Diskussion des Begriffs der Dimensionalität vgl. Böhme & Robitzsch (2009a).

2.5 Sprache als Gegenstand wissenschaftlicher Forschung

Sprache ist der Untersuchungsgegenstand verschiedenster wissenschaftlicher Disziplinen, einige von diesen sollen nachfolgend kurz dargestellt werden.

Die wissenschaftliche Erforschung von Sprache ist in erster Linie Gegenstand der *Linguistik* beziehungsweise der Sprachwissenschaft. Diese versucht die Beschaffenheit von Sprache als System kennzeichnend zu beschreiben und sprachliche Phänomene zu erklären. Hierbei interessiert sie sich nicht primär für Sprache als Werkzeug menschlicher Kommunikation und Interaktion. Vielmehr konzentriert sich die Linguistik auf die sprachliche Regelhaftigkeit und Produktivität und betrachtet Sprache als ein System komplexer Strukturen. Zur Erklärung der Regelhaftigkeit bezieht sich die Linguistik auf eine Grammatik, die es gestattet, zur Phonologie (den Sprachlauten), zur Syntax (der Sprachstruktur) sowie zur Semantik (der Bedeutung) Aussagen zu treffen. Ferner behandelt die Linguistik die Struktur von Mitteilungen und beschreibt diese als ein objektives Gebilde, dessen Organisationsregeln es zu erkennen gilt.

Die *Sprachpsychologie* oder Psycholinguistik hingegen zielt auf eine wissenschaftliche Beschreibung von Sprachproduktion und Sprachrezeption sowie deren Zusammenspiel im Kommunikationsgeschehen. Hierbei wird unter dem Begriff der *Sprachproduktion* die Erzeugung von geordneten Sprachlautsequenzen sowie die vorhergehenden und begleitenden psychischen Prozesse der Planung beziehungsweise Kontrolle und Regulation verstanden (vgl. Herrmann, 1992). Auch die Erzeugung nicht-phonetischer Sprachprodukte, also das Schreiben, zählen zur Sprachproduktion, stehen jedoch selten im Mittelpunkt sprachpsychologischer Betrachtungen. Die Wahrnehmung, also das Hören des vom Partner Gesprochenen und die entsprechenden psychischen Begleitprozesse werden als Sprachverstehen oder *Sprachrezeption* bezeichnet. Wiederum zählt auch das Lesen von Geschriebenem beziehungsweise Gedrucktem zum Bereich der Sprachrezeption, wobei dieser Aspekt in der Sprachpsychologie ebenfalls nur wenig Aufmerksamkeit erfährt. Die Prozesse der Sprachproduktion und Sprachrezeption gelten als eng miteinander verzahnt, da Menschen in aller Regel in Kommunikationssituationen als „integrierte Hörer/Sprecher“ (Herrmann, 1992, S. 283) verstanden werden.

Kommunikationssituationen werden in der Sprachpsychologie in einer ersten Näherung als Informationsverarbeitungsprozess konzipiert, bei dem Vorgänge des Enkodierens und Dekodierens sowie Zustände von Mitteilungen mit Zuständen von Sendern und Empfängern in Beziehung gesetzt werden. Legt man diese Einbettung zugrunde, ereignet sich sprachliche Kommunikation in einem sozialen Feld, das durch die beiden Pole Sender und Empfänger (bzw. oftmals Sprecher und Hörer) bestimmt wird. Durch eine Mitteilung sind Sprecher und Hörer miteinander verbunden, wobei die Mitteilung oder Nachricht gleichzeitig den Output des

Sprechers sowie den Input des Hörers darstellt. Die Tätigkeit des Sprechers, durch welche die Mitteilung aufgebaut und übermittelt wird, wird hierbei als Enkodieren bezeichnet. Die Tätigkeit des Hörers, die die ankommenden Schallwellen in Bedeutung umwandelt, wird als Dekodieren charakterisiert. Enkodieren und Dekodieren sind somit Übersetzungsvorgänge von Bedeutung in Sprache und von Sprache in Bedeutung. Begrenzt wird die Übersetzung durch die Möglichkeiten, welche Lexikon und Grammatik bereitstellen.

Sowohl Sprachproduktion als auch Sprachrezeption gelten in der Sprachpsychologie als wesentliche Voraussetzungen, um sich in verschiedenen Situationen zurechtzufinden und diese durch Handeln bewältigen zu können. Hierbei dienen sowohl Sprachverstehen als auch (die mündliche) Sprachproduktion der Reduktion von Soll-Ist-Differenzen und somit der Handlungsregulation (vgl. Herrmann, 1992, S. 287).

Ferner ist Sprache als soziales Phänomen Gegenstand sprachpsychologischer Betrachtungen. Sprachliche Kommunikation begleitet und ermöglicht interindividuelle Interaktion. Vor allem in Bezug auf die Kommunikation komplexer Sachverhalte ist Sprache unerlässlich, da hier nonverbale Kommunikation oftmals nicht genügt.

Neben der Linguistik und der Sprachpsychologie ist Sprache auch in zahlreichen anderen Teildisziplinen ein zentraler Gegenstand, wobei je nach Disziplin unterschiedliche Blickwinkel vorherrschen und verschiedene wissenschaftliche Forschungsmethoden zum Einsatz kommen. So liefert beispielsweise die Entwicklungspsychologie wichtige Erkenntnisse zu den affektiven und kognitiven Aspekten bei der Bewältigung der Entwicklungsaufgabe des Spracherwerbs. Wesentliche Stationen des frühen Sprachlernprozesses und seine Implikationen für die vorliegende Arbeit werden im nachfolgenden Abschnitt 2.6 dargestellt.

Was dieser kurze Abschnitt gezeigt hat, ist, dass die wissenschaftliche Untersuchung von Sprache Gegenstand zahlreicher verschiedener Disziplinen ist. So zahlreich wie die beteiligten Wissenschaften sind die hierbei gewählten Perspektiven, die untersuchten Fragestellungen und das eingesetzte Methodenrepertoire. Nicht alle vorherrschenden wissenschaftlichen Sichtweisen konnten an dieser Stelle vorgestellt und erläutert werden. Die Darstellung verbleibt somit notwendigerweise exemplarisch. Was dennoch deutlich geworden sein sollte: Eine Arbeit wie die vorliegende kann sich bei der empirischen Erforschung sprachlicher Kompetenzen zwar um die Berücksichtigung interdisziplinärer Vielfalt bemühen, dennoch wird das Resultat stets ein eingeschränkter Blick auf die untersuchten Phänomene sein, da hinsichtlich des wissenschaftlichen Selbstverständnisses, der untersuchten Fragen und der hierfür gewählten Methoden Entscheidungen getroffen werden müssen, die alternative Zugänge ausschließen.

2.6 Exkurs: Sprachentwicklung

Bevor die Bildungsstandards für das Fach Deutsch im Primarbereich (KMK, 2005a) vertieft betrachtet werden, soll vorab die frühe Entwicklung der Sprachkompetenz vor dem Beginn der Schulzeit skizziert werden. Diese Phase der Entwicklung und Ausdifferenzierung sprachlicher Kompetenzen ist für diese Dissertation insoweit von Bedeutung, als sie einige wichtige Hinweise für die Interpretation empirischer Befunde liefern kann.

Das sprachliche Wissen eines Kindes steht in enger Beziehung zu seinen kognitiven und sozialen Fähigkeiten sowie seiner Verhaltensregulation und gestattet Vorhersagen über seine späteren Entwicklungsmöglichkeiten (vgl. Grimm & Wilde, 1998). Hierbei ist auch der frühe Spracherwerb wesentlich durch das Bedürfnis nach sozialer Interaktion geprägt. Dies wird später näher erläutert, zunächst soll der Spracherwerb mit seinen wesentlichen Schritten kurz skizziert werden.

Im Verlauf der Sprachentwicklung erwerben Kinder Fähigkeiten auf drei unterschiedlichen Ebenen. Zunächst werden prosodische Kompetenzbestandteile erworben, die über die Rhythmik von Spracheinheiten definiert sind und an denen Tonhöhe, Lautstärke, die Länge der Sprachlaute sowie die Pauseneinteilung beteiligt sind. In einem zweiten Schritt eignen sich die Kinder linguistische Kompetenzbestandteile an, welche neben Phonologie, Morphologie und Syntax auch das Lexikon, also Wörter und ihre Bedeutungen, umfasst. Auf einer abschließenden dritten Ebene werden schließlich pragmatische Kompetenzbestandteile erworben, welche sich durch die angemessene kommunikative Verwendungsweise von Sprache je nach Kontext auszeichnet.

Der frühe Spracherwerb umfasst von den ersten Sprachlauten bis hin zur Wortproduktion fünf wesentliche Schritte (vgl. bspw. Grimm & Wilde, 1998). Bereits wenige Wochen nach der Geburt lassen sich bei Säuglingen erste, zunächst rein akustische Lautbildungen beobachten, die im Alter von etwa zwei bis drei Monaten in ein Gurrstadium übergehen. In dieser Phase werden bereits sprachliche Laute produziert und vorgesprochene Vokale nachgeahmt. In der nachfolgenden Phase der Expansion nehmen die Gurrlaute im Alter von etwa vier bis fünf Monaten zunehmend Ähnlichkeit mit realen Sprachlauten an. Das so genannte kanonische Lallstadium wird im Alter von sechs bis neun Monaten erreicht. In dieser Phase reduplizieren die Säuglinge Konsonant-Vokal-Verbindungen und verwenden hierbei wort- und satzähnliche Intonation. Schließlich mündet die phonologische Entwicklung im Alter von etwa zehn bis vierzehn Monaten in der Produktion der ersten Wörter. An diese frühen Phasen des Spracherwerbs schließt sich die lexikalische Entwicklung an. Im Stadium des zu Beginn vorherrschenden langsamen Wortlernens bilden vermutlich assoziative Prozesse im Sinne eines

Paar-Assoziations-Lernens in einem sozial-interaktiven Lernkontext den Hintergrund des Worterwerbs. Als entwicklungskritisch wird das Erreichen der 50-Wörter-Grenze im Alter von ungefähr 18 Monaten angesehen. Anschließend werden sehr schnell viele neue Wörter dazugelernt, weshalb diese Phase mitunter als „Benennungsexplosion" bezeichnet wird. Dieses schnelle Wortlernen für Objekte und Objektmerkmale zeigt eine neue Qualität, da die Kinder nun sehr schnelle Zuordnungen zwischen neuen Wörtern und meist noch unvollständigen Bedeutungen vornehmen (*fast mapping*). Ausdruck dieser noch unvollständigen Bedeutungsrepräsentationen sind Übergeneralisierungen sowie Überdiskriminierungen. Die Lernmechanismen nach dem Erreichen der 50-Wörter-Grenze besitzen eine neue Qualität und werden vermutlich durch so genannte *constraints* gesteuert, welche die möglichen Bedeutungen von Wörtern einschränken und auf diese Weise dazu dienen, sehr schnell eine Verknüpfung von Wort und (wahrscheinlicher oder ungefährer) Bedeutung herzustellen. Erst später, in der Phase der Wortexplosion, werden die Voraussetzungen für den Grammatikerwerb geschaffen, da während dieser Phase neben Nomen auch zunehmend Verben erlernt werden. Eine kritische Größe des Verbwortschatzes wiederum ist die Vorbedingung für das Ableiten grammatischer Regeln.

Kinder bewältigen diese Phasen des Spracherwerbs in einem Alter, in dem sie zu vergleichbar anspruchsvollen Ergebnissen in anderen kognitiven Bereichen noch nicht befähigt sind. Um diese enorme Leistung deuten zu können, genügen rein linguistische Erklärungsmodelle nicht aus. Aktuellere Ansätze gehen davon aus, dass eine Universalgrammatik im Sinne Chomskys nicht Voraussetzung für den Spracherwerb ist, sondern sein Ergebnis. Es wird diskutiert, dass Kinder sich in ihrer Sprachentwicklung von der kommunikativen Funktion hin zur Struktur entwickeln – und nicht umgekehrt. Als wesentliche angeborene Komponenten gelten allerdings dennoch die folgenden drei Aspekte:

- phonologische Wahrnehmungsfähigkeiten vor und kurz nach der Geburt,
- verschiedene affektive und kognitive Voraussetzungen, die unter anderem das Bedürfnis nach Kommunikation und die Differenzierungs- und Kategorisierungsfähigkeit einschließen und
- für den Erwerb kritische Zeitfenster.

Man geht davon aus, dass diese angeborenen Komponenten sich allerdings erst in der sozialen Interaktion entfalten können. Wesentliche Bedeutung kommt somit dem Zusammenwirken von rein sprachlichen mit sozial-affektiven sowie kognitiven Mechanismen zu. Die Bedeutung der sozial-interaktiven Komponente in der beginnenden sprachlichen Kommunikation unterstreichen

Grimm und Wilde (1998) mit der folgenden Äußerung: „Das Kind äußert nicht akustisch wahrnehmbare Lauthülsen ohne affektiven oder kognitiven Gehalt. Es äußert auch keine Wörter oder Sätze im sozialen Vakuum, sondern in der sozialen Interaktion“ (S. 452).

Für die vorliegende Arbeit ist die sprachliche Entwicklung des Menschen insofern relevant, als deutlich geworden sein sollte, dass der Spracherwerb einen hochkomplexen Prozess darstellt, der sich über einen langen Zeitraum erstreckt. Während der Sprachentwicklung werden die sprachlichen Teilfähigkeiten nicht simultan, sondern zeitlich aufeinander folgend erlernt. Dies bedeutet, dass das (Zu-)Hören die erste sprachliche Fähigkeit des Menschen darstellt. Erst deutlich später erlernt das Kind sukzessive das Sprechen. Beide mündlichen Sprachfähigkeiten liegen hinsichtlich ihres Erwerbszeitraums wiederum beträchtlich vor dem Erlernen der schriftsprachlichen Fähigkeiten, welche traditionell zu Beginn der Primarphase unterrichtet werden. Nach wie vor gilt es als ureigene Aufgabe der Beschulung in der Primarstufe, allen Kindern das Lesen und Schreiben beizubringen (vgl. Büker & Vorst, 2010). Allerdings greifen die Kinder zu Schulbeginn heutzutage oftmals auf sehr verschiedene schriftsprachliche Erfahrungsschätze zurück und zeigen bei Schuleintritt große Differenzen in der Leseentwicklung. Der Leseanfangsunterricht trifft somit auf „ausgesprochen heterogene Lernvoraussetzungen“ (Büker & Vorst, 2010, S. 21). Es kann also nicht mehr als allgemeine Zielstellung gelten, zunächst als elementare Grundlage die Beherrschung des Schriftsystems als Voraussetzung der Lesefertigkeit im Sinne des Dekodierens zu vermitteln, um anschließend das Leseverstehen zu fördern. Vielmehr rücken aufgrund der heterogenen Lernvoraussetzungen zunehmend individualisierte Unterrichtskonzepte in den Vordergrund, die eine optimale Förderung entsprechend der verschiedenen Kompetenzstände ermöglichen (vgl. Büker & Vorst, 2010).

3 Einführung der in den Bildungsstandards thematisierten sprachlichen Kompetenzen

3.1 Die Struktur sprachlicher Kompetenzen[7]

Für die Messung sprachlicher Kompetenzen ist die Frage zentral, ob ein übergreifender Faktor sprachlicher Fähigkeit existiert, auf den sämtliche sprachliche Leistungen zurückgeführt werden können, oder ob verschiedene sprachliche Leistungen als ein mehrdimensionales Konstrukt verstanden werden sollten. Hierbei muss eine theoretisch-deskriptive Konzeption der Struktur sprachlicher Kompetenzen von der empirischen Prüfung der Konstruktdimensionalität differenziert werden. Während es beispielsweise aus Sicht der Orthografietheorie sinnvoll ist, in Anlehnung an die Entwicklung der orthografischen Kompetenz eine Vielzahl orthografischer Teilkompetenzen zu beschreiben, muss sich diese deskriptiv gelingende Differenzierung nicht zwangsläufig in empirischen Befunden in dem Sinne niederschlagen, dass mehrere weitgehend voneinander unabhängige Teildimensionen auch datenanalytisch unterschieden werden können (vgl. Böhme & Bremerich-Vos, 2009). Nichtsdestotrotz sind sowohl die theoretische Betrachtung als auch die empirische Prüfung der Struktur sprachlicher Kompetenzen für die testdiagnostische Umsetzung der Bildungsstandards im Fach Deutsch relevant. Je nachdem, ob von einer allgemeinen sprachlichen Kompetenz oder von relativ unabhängigen Teilkompetenzen auszugehen ist, hat dies Auswirkungen auf Aufbau und Inhalt der eingesetzten Testinstrumente (vgl. hierzu Böhme, Neumann & Schipolowski, 2010; Jude, 2008).

Die pädagogisch-psychologische Diagnostik betreibt Strukturanalysen, also die Untersuchung der *empirischen* Trennbarkeit verschiedener Konstrukte, mit Hilfe der Erfassung von Unterschieden in Fähigkeitsausprägungen und somit aus einer differentiellen Perspektive. Nur selten basieren Analysen zur Struktur kognitiver Kompetenzen auf längsschnittlichen Daten und somit auf einem entwicklungspsychologischen Ansatz – wobei die Vernachlässigung dieser Perspektive in erster Linie auf die wesentlich aufwändigeren Datenerhebungen und nicht auf ihre mangelnde Eignung zurückzuführen ist.

---

[7] Der Abschnitt „3.1 Die Struktur sprachlicher Kompetenzen" basiert teilweise auf Textpassagen des Beitrags von Bremerich-Vos, A., Böhme, K. & Robitzsch, A. (2009). Sprachliche Kompetenzen im Fach Deutsch – Strukturanalysen und Validierungsbefunde. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 198-218). Weinheim: Beltz.

In Abhängigkeit vom gewählten Blickwinkel können zahlreiche verschiedene Strukturkomponenten der Sprachkompetenz im Fokus der Aufmerksamkeit stehen. Denkbar sind die Differenzierung von Erst- und Fremdsprachen, die Unterscheidung von produktiven und rezeptiven Sprachkompetenzen, die Differenzierung von mündlichen und schriftsprachlichen Kompetenzen oder die Unterscheidung von Kompetenzbereichen, wie sie im Rahmen der Bildungsstandards für das Fach Deutsch im Primarbereich und der Sekundarstufe I (vgl. KMK, 2004, 2005a, 2005b) vorgenommen wurden. Hier werden die Bereiche 1) *Sprechen und Zuhören*, 2) *Schreiben*, 3) *Lesen – mit Texten und Medien umgehen* sowie 4) *Sprache und Sprachgebrauch untersuchen* unterschieden, wobei der vierte Bereich „in Beziehung zu jedem der drei anderen Bereiche" steht (KMK, 2004, S. 7). Die Deutschdidaktik gebraucht in diesem Kontext allerdings selten den Begriff Kompetenzbereich, sondern verwendet primär den Terminus *Arbeitsbereich*. Prominent, wenn auch teils kritisch diskutiert (vgl. Steinbrenner, 2007), ist hier die Unterscheidung von Ossner (2006a), der die Bereiche 1) *Sprechen und Zuhören*, 2) *Schreiben*, 3) *Lesen und Verstehen* sowie 4) *Sprache thematisieren* differenziert. Wesentlicher Unterschied zur Strukturierung der Bildungsstandards besteht darin, dass die ersten drei Bereiche von Ossner (2006a) in den Konzepten von Mündlichkeit und Schriftlichkeit (vgl. P. Koch & Österreicher, 1985) verankert werden, der Arbeitsbereich *Sprache thematisieren* aber eigenständig besteht.

Während die Frage nach der theoretischen wie empirischen Struktur muttersprachlicher Kompetenzen im deutschen Sprachraum bislang nur wenig Aufmerksamkeit erfahren hat, wurde dieses Thema im Rahmen der englischsprachigen Literatur zum *Language Testing* breit diskutiert. In den 70er Jahren des vergangenen Jahrhunderts äußerte Oller seine prominente These, dass Sprache eine einheitliche Fähigkeit darstellt (Oller, 1976). Auch in der Sprachpsychologie gelten Sprachproduktion und Sprachrezeption als untrennbar miteinander verknüpft (vgl. Herrmann, 1992), ihre Entwicklung erfolgt in enger Wechselwirkung (vgl. Abschnitt 2.6) und auch die ihnen zugrunde liegenden Hirnstrukturen sind benachbart und in komplexer Weise vernetzt (vgl. Friederici, 1984). Wird Sprache als soziales Phänomen verstanden, bilden Sprachproduktion und Sprachrezeption die stets aufeinander bezogenen Kernelemente menschlicher Kommunikation und insofern eine Einheit. Offen bleibt dennoch, welche Binnenstruktur sprachlicher Kompetenzen theoretisch denkbar wäre und empirisch ermittelt werden kann.

In der deutschdidaktischen Literatur ist bislang kaum erörtert worden, welche Struktur die (mutter- oder verkehrs-)sprachlichen Kompetenzen von Schülerinnen und Schülern haben. Darüber hinaus sind die wenigen vorliegenden Beiträge überwiegend theoretischer Natur (vgl. bspw. Felder, 2003). Bislang gilt als Konsens, dass jeweils im Medium des Mündlichen und des

Schriftlichen eher produktive und eher rezeptive Kompetenzen unterschieden werden können (vgl. Tabelle 2).

Im Hinblick auf die in den Bildungsstandards unterschiedenen Kompetenzbereiche bleibt zunächst ungeklärt, wo in diesem Bezugssystem der Bereich *Sprache und Sprachgebrauch untersuchen* zu verorten ist (vgl. hierzu auch Ossner, 2006a). Die implizite oder auch explizite Verfügbarkeit von sprachlichen Mitteln kann als Stützfunktion für die rezeptiven und produktiven sprachlichen Kompetenzen verstanden werden, wie dies beispielsweise in Modellen für fremdsprachliche Kompetenzen der Fall ist (vgl. Europarat, 2001).

Tabelle 2: Strukturierung sprachlicher Kompetenzen

	Mündlich	**Schriftlich**
Sprachproduktion	Sprechen	Schreiben/Rechtschreibung
Sprachrezeption	Zuhören	Lesen

Anders als in der Deutschdidaktik hat die empirische Analyse der Struktur sprachlicher Kompetenzen im Rahmen der psychologischen Diagnostik und auch der empirischen Bildungsforschung eine lange Tradition. Allerdings beziehen sich die einschlägigen Studien zumeist auf das Englische und dort oftmals auf Englisch als Zweit- beziehungsweise Fremdsprache (vgl. bspw. Köller & Trautwein, 2004; Leucht, Retelsdorf, Möller & Köller, 2010; Sang, Schmitz, Vollmer, Baumert & Roeder, 1986; Sawaki, Stricker & Oranje, 2009; Shin, 2005; Song, 2008), so dass die Übertragbarkeit der Befunde auf die Struktur des Deutschen als Mutter- beziehungsweise Verkehrssprache fraglich bleibt.

Strukturuntersuchungen des Englischen basieren oftmals auf Analysen zum *Test of English as a Foreign Language* (TOEFL). Dieser Test wurde in den 1960er Jahren vom *Educational Testing Service* (ETS) entwickelt und dient bis heute der Überprüfung der Sprachfertigkeiten von internationalen Studienplatzbewerbern, die ein Studium an einer nordamerikanischen Universität aufnehmen möchten. Ursprünglich als *Paper-Pencil-Test* konzipiert, liegen mittlerweile auch computerbasierte sowie internetbasierte Versionen vor, die neben dem Lese- und Hörverstehen auch die Schreib- und Sprechkompetenz überprüfen (vgl. Sawaki, Stricker & Oranje, 2009).

Frühe Untersuchungen für das Englische als Zweit- beziehungsweise Fremdsprache führte Oller (1976) unter Rückgriff auf Hauptkomponentenanalysen durch. Er interpretierte seine Befunde für verschiedene Sprachtests, unter anderem für den TOEFL, so, dass den Leistungen

der Probanden jeweils ein einziger Generalfaktor (*g*-Faktor) zugrunde liegen müsse. In der einschlägigen Literatur wird diese Annahme unter dem Begriff der *Unitary Competence Hypothesis* diskutiert. Das Vorgehen von Oller wurde jedoch aus methodischer Sicht verschiedentlich kritisiert (vgl. Vollmer & Sang, 1983; Sang et al., 1986). So wurde gezeigt, dass das Generalfaktormodell über verschiedene Stichproben oder verschiedene analytische Zugänge hinweg nicht invariant blieb (Sang et al., 1986).

Köller und Trautwein (2004) untersuchten die dimensionale Struktur des TOEFL im Rahmen der TOSCA-Studie[8]. Neben der Überprüfung der rezeptiven Fremdsprachenkompetenz in zwei Untertests, je einem für das Hör- und Leseverstehen, wurden hier in einem dritten Untertest die Grammatik- und Orthografiekenntnisse ermittelt. Letztere werden in der Fremdsprachendidaktik häufig unter dem Begriff der Sprachbewusstheit zusammengefasst. Die Autoren nahmen in ihrer Studie die Strukturanalysen mittels konfirmatorischer Mehrebenen-Faktorenanalysen vor und kamen zu dem Ergebnis, dass ein Modell, welches auf Schulebene einen *g*-Faktor und auf Individualebene drei korrelierte Faktoren postuliert, eine ähnlich gute Passung aufweist wie ein Modell mit einem *g*-Faktor auf Schulebene und einem *g*-Faktor sowie drei unkorrelierten Faktoren auf Individualebene. Unter Berücksichtigung des Sparsamkeitskriteriums bevorzugen Köller und Trautwein (2004) das erstgenannte Modell. Insgesamt schlussfolgern die Autoren, dass die Verwendung eines Gesamttestwerts auf Schulebene gerechtfertigt scheint. Ergänzend weisen die Autoren allerdings in einer Fußnote darauf hin, dass ein Modell, welches sowohl auf Schul- als auch auf Individualebene nur jeweils einen *g*-Faktor annimmt, keine zufriedenstellenden Gütekriterien aufweist und den differenzierteren Modellen somit unterlegen ist. Dementsprechend scheint eine Beschränkung auf die Modellierung eines Generalfaktors – insbesondere auf Individualebene – der tatsächlichen Konstruktstruktur nicht gerecht zu werden. Neben einem Generalfaktor sind somit zusätzlich spezifische Faktoren zu berücksichtigen. Eben dies wird auch in der Rückmeldung der Testergebnisse umgesetzt, bei der neben dem Gesamttestwert ergänzend Leistungsresultate für die Untertests mitgeteilt werden.

Hinsichtlich einer neuentwickelten internetbasierten Version des TOEFL, die Aufgaben zur Lese-, Zuhör-, Sprech- und Schreibkompetenz umfasst (TOEFL iBT), berichten Sawaki, Stricker und Oranje (2009): "The Higher-Order Factor model that included a single higher-order general factor (ESL/EFL ability) and four first-order factors corresponding to the four TOEFL iBT sections (modalities) was the best representation of the test's factor structure" (S. 24). Dieser

[8] Das Akronym TOSCA steht für „Transformation des Sekundarschulsystems und akademische Karrieren" (Köller, Watermann, Trautwein & Lüdtke, 2004).

Befund stimmt mit der derzeitigen Sichtweise der Struktur sprachlicher Kompetenzen für das Englische als Zweit- beziehungsweise Fremdsprache sehr gut überein, welche sowohl einen Generalfaktor als auch spezifische Faktoren annimmt (vgl. Sawaki, Stricker & Oranje, 2009).

Andere Studien zur Dimensionalität der sprachlichen Kompetenzen beziehen sich – wiederum hauptsächlich für den zweit- oder fremdsprachlichen Bereich – auf die Frage, ob die Struktur sprachlicher Kompetenzen vom jeweiligen Fähigkeitsniveau abhängt. Wiederum Bezug nehmend auf Untersuchungen des TOEFL fanden Ginther und Stevens (1998) beispielsweise für Spanisch als Fremdsprache Hinweise darauf, dass die Teilkompetenzen Zuhören, Lesen, Schreiben und Sprechen mit zunehmender Fähigkeit voneinander unabhängiger werden. Auch Shin (2005) interessierte sich dafür, ob bei höherem Fähigkeitsniveau mit eher weniger oder eher mehr unterscheidbaren Subdimensionen zu rechnen ist. Die Studie bezog sich auf Daten von 779 Probanden, die den TOEFL sowie SPEAK (*Speaking Proficiency in English Assessment Kit*) bearbeitetet hatten. Mittels Multigruppen-Strukturgleichungsmodellen prüfte Shin (2005) ein Generalfaktormodell, Modelle höherer Ordnung und Modelle korrelierter Faktoren, konnte aber keine eindeutigen und verallgemeinerbaren Aussagen zur Invarianz der Struktur sprachlicher Kompetenzen über verschiedene Fähigkeitsniveaus hinweg treffen.

Im deutschen Sprachraum untersuchten Leucht, Retelsdorf, Möller und Köller (2010) mittels bildungsstandardbasierter Testaufgaben den Zusammenhang zwischen den beiden rezeptiven Sprachkompetenzen für das Englische als Fremdsprache. Hierbei überprüften sie die Lese- und Hörverstehensleistungen von Schülerinnen und Schülern der achten Jahrgangsstufe und kamen zu dem Ergebnis, dass die beiden rezeptiven Kompetenzen sehr hoch miteinander assoziiert sind. Die Prüfung der Zusammenhänge mit Drittvariablen fiel für das Hör- und Leseverstehen insgesamt sehr ähnlich aus, was gegen eine differenzielle Validität der zwei rezeptiven Kompetenzen spricht und ebenfalls die Interpretation nahelegt, dass ein Generalfaktor für das Textverstehen angenommen werden kann.

Für Deutsch als Fremdsprache (DaF) berichten Eckes und Grotjahn (2006) im Zusammenhang mit Validitätsanalysen für C-Tests[9] über die Strukturen sprachlicher Kompetenzen im Deutschen. Diese Strukturprüfungen beziehen sich auf den TestDaF[10] und C-Tests, die von 843 Teilnehmern bearbeitet wurden. In konfirmatorischen Faktorenanalysen prüften die Autoren ein Globalfaktormodell, in dem Testergebnisse zum Lesen, Hören, Schreiben, Sprechen sowie bei C-Tests Indikatoren einer allgemeinen Sprachkompetenz sind. Ferner wurden Varianten zweifaktorieller Modelle geprüft, die sich entweder auf die

[9] C-Tests sind schriftliche Tests der allgemeinen Sprachbeherrschung in der Muttersprache oder einer Fremdsprache, die auf dem Prinzip der reduzierten Redundanz beruhen (vgl. Grotjahn, 2002).

[10] TestDaF steht für Test Deutsch als Fremdsprache (vgl. http://www.testdaf.de)

Unterscheidung von Rezeption und Produktion oder auf die von Mündlichkeit und Schriftlichkeit beziehen. Zentraler Befund von Eckes und Grotjahn (2006) ist, dass die Passung der zweifaktoriellen Modelle kaum besser als die eines Globalfaktormodells ist. Die Autoren schlussfolgern: „In view of the consistently high factor correlations and the negligible differences in fit between the one-factor model and the two-factor models, it seems reasonable to prefer the more parsimonious one-factor model" (Eckes & Grotjahn, 2006, S. 315). Da allerdings einige Unstimmigkeiten innerhalb einer Teilstichprobe auftraten und der Stichprobenumfang insgesamt eher gering war, mahnen die Autoren zur Vorsicht und warnen vor einer unreflektierten Verallgemeinerung dieses Befundes.

Im Rahmen der DESI-Studie (Deutsch Englisch Schülerleistungen International; Beck & Klieme, 2007; DESI-Konsortium, 2008) wurden die Kompetenzen im Schreiben beziehungsweise der Textproduktion, im Rechtschreiben, Lesen, der Sprachbewusstheit und darüber hinaus in Bezug auf Wortschatz und Argumentation geprüft. Ein Modell, wonach für Leistungsvariationen in all diesen sprachlichen Kompetenzen ein einziger Generalfaktor verantwortlich ist, ergab eine nur befriedigende Passung. Bessere Werte resultierten für ein Mehrebenenmodell, in dessen Rahmen auf der Individualebene eine mehrdimensionale Struktur und auf der Ebene der Klasse ein Modell mit zwei latenten Faktoren – Reflexion/Rezeption und Produktion – angenommen wurde. Die Korrelationen der genannten Subdimensionen liegen auf Individualebene zwischen $r = .15$ und $r = .36$, wobei die Zusammenhänge zwischen Lesen und Rechtschreiben am geringsten und die zwischen Wortschatz und Sprachbewusstheit sowie zwischen Rechtschreibung und dem Aspekt der Textproduktion, der sich auf Richtigkeit bezieht, am höchsten sind (Jude et al., 2008, S. 195). Die Größenordnung der auf Individualebene gefundenen Korrelationen zwischen den verschiedenen Teilbereichen erscheint insgesamt überraschend gering. Bremerich-Vos, Böhme und Robitzsch (2009) setzen sich daher vertieft mit den Befunden von Jude und Kollegen (2008) auseinander und verdeutlichen, dass die geringe Höhe der Zusammenhänge ein Resultat des mehrebenenanalytischen Vorgehens ist. Neben diesem methodischen Aspekt ist inhaltlich nicht unmittelbar einsichtig, dass auf der Ebene der Klasse neben dem Produktionsfaktor (Schreiben und Rechtschreiben) in einem weiteren Faktor Reflexion und Rezeption zusammengefasst sind. Auf diesem zweiten Faktor laden neben Aufgaben zum Wortschatz auch Aufgaben zur Sprachbewusstheit und zum Argumentieren, bei denen durchaus eine Produktionsleistung (z. B. von Wörtern und Argumenten) zu erbringen war. Aufgaben zur Überprüfung der Kompetenzstände im Bereich der Mündlichkeit kamen im Rahmen der DESI-Studie nur für das Englische, nicht aber für das Deutsche vor. Zwar sollten die Schülerinnen und Schüler im Bereich der Argumentation Gespräche analysieren, diese wurden aber nur in schriftlicher Form vorgelegt. Insofern konnte anhand der in DESI

verfügbaren Daten zwar geprüft werden, ob sich die Dimensionen Produktion und Rezeption voneinander unterscheiden lassen. Inwieweit auch das Medium, also die Unterscheidung von Mündlichkeit vs. Schriftlichkeit, für die dimensionale Struktur von Bedeutung ist, konnte im Rahmen von DESI hingegen nicht empirisch untersucht werden.

Jude (2008), die sich im Kontext ihrer Dissertation intensiv mit Strukturanalysen sprachlicher Kompetenzen im Deutschen und Englischen anhand der DESI-Daten auseinandersetzt, kommt zu dem folgenden Resultat:

> Das grundlegende Ergebnis der vorliegenden Arbeit ist die Aussage, dass *Sprachkompetenz als mehrdimensionales Konstrukt* anzusehen ist. […] Im Vergleich unterschiedlicher Modelle erweist sich das eindimensionale Modell, das jegliche Varianz in den Teilbereichen auf einen einzigen gemeinsamen Faktor zurückführt, als unzureichend. Vielmehr ist zu beobachten, dass mehrere latente Dimensionen von Sprachkompetenz existieren, die jeweils spezifische Teilbereiche auf sich vereinen und darüber hinaus systematisch miteinander korrelieren. (Hervorhebungen im Original, Jude, 2008, S. 181)

Insgesamt verdeutlichen die bislang berichteten Befunde zur Struktur sprachlicher Kompetenzen, dass sich nur schwer durchweg verallgemeinerbare Ergebnisse identifizieren und somit kaum allgemeingültige Aussagen treffen lassen. Dennoch zeichnet sich ab, dass ein Generalfaktor allein oftmals nicht ausreicht, um die Struktur sprachlicher Kompetenzen erschöpfend abzubilden. Vielmehr scheint es angezeigt, ergänzend spezifische Faktoren für die unterschiedlichen sprachlichen Teilbereiche anzunehmen.

An dieser Stelle ergibt sich die Frage, wie sich die empirische Befundlage hinsichtlich der Struktur der in den Bildungsstandards für das Fach Deutsch im Primarbereich (KMK, 2005a) beschriebenen sprachlichen Kompetenzen gestaltet. Entsprechende Analysen legten Bremerich-Vos, Böhme und Robitzsch (2009) vor. Untersucht wurden die Beziehungen der Teilfähigkeiten des Zuhörens, Lesens, Schreibens, der Orthografie und der Sprachreflexion. Paarweise zweidimensionale Analysen der genannten Bereiche ergaben besonders hohe Korrelationen zwischen den Kompetenzbereichen Zuhören und Lesen, darüber hinaus zwischen der Schreib- und der Rechtschreibkompetenz. Das spricht zunächst für die Annahme, dass sich produktive und rezeptive Teilfähigkeiten theoriekonform voneinander separieren lassen. Eine Hauptkomponentenanalyse der Teilkompetenzen Zuhören, Schreiben, Lesen, Sprachgebrauch, Orthografie sowie einer Skala zur Prüfung der verbalen Intelligenz ergab einen dominanten ersten Faktor. Dieser klärte mit 66,4 Prozent einen erheblichen Teil der Varianz auf und kann als Globalfaktor sprachlicher Kompetenz interpretiert werden (Bremerich-Vos, Böhme & Robitzsch, 2009, S. 211). Entsprechend der theoretisch plausiblen Unterscheidung von produktiven und

rezeptiven Sprachkompetenzen wurde ergänzend eine explorative Faktorenanalyse mit zwei Faktoren berechnet. Die Befunde ergaben plausible Anordnungen der verschiedenen Kompetenzbereiche entsprechend der Unterscheidung von produktiven und rezeptiven Fähigkeiten, wobei allerdings keine vollständige Trennung der entsprechenden Kompetenzen erzielt werden konnte. Erwartungskonform zeigten Schreiben, Orthografie und die zusätzlich erhobene verbale Intelligenz hohe Ladungen auf einem Faktor, der als „produktiv" interpretiert werden kann. Für den rezeptiven Faktor zeigt das Zuhören eine sehr hohe Ladung. Die Kompetenz im Bereich „Sprache und Sprachgebrauch" und vor allem die Lesekompetenz sind aber nicht eindeutig zuzuordnen und weisen somit Merkmale beider Dimensionen auf. Damit wird zum einen der integrative Charakter der Kompetenz im Bereich der Sprachreflexion unterstrichen. Zum anderen wird deutlich, dass der rezeptive Charakter des Lesens durch zahlreiche in der Testung schriftlich-produktiv zu erledigende Arbeitsaufträge konfundiert und nicht separat betrachtbar ist. Insgesamt kann dieses Befundmuster so verstanden werden, dass das Hörverstehen eine sehr spezielle Teilfähigkeit innerhalb der sprachlichen Kompetenzen im Deutschen darstellt und dass die Trennung in rezeptive und produktive Fähigkeiten nicht durchgehend überzeugend gelingt (Bremerich-Vos, Böhme & Robitzsch, 2009, S. 212 f.).

Im Rahmen der Validierung an Außenkriterien ergaben sich erwartungskonforme latente Korrelationen zwischen den Aufgaben zur Überprüfung der Erreichung der Bildungsstandards mit Aufgaben aus dem Itempool der Vergleichsarbeiten in der dritten Jahrgangsstufe (VERA-3 Deutsch) in Höhe von $r = .74$ sowie mit dem Lesetest HAMLET 3-4 (Lehmann, Peek & Poerschke, 2006) in Höhe von $r = .84$ (Bremerich-Vos, Böhme & Robitzsch, 2009, S. 213). Die Prüfung der Zusammenhänge mit Noten in den Fächern Deutsch und Mathematik fiel ebenfalls erwartungskonform aus. Die Zusammenhänge der Testleistungen in den sprachlichen Kompetenzen mit der Deutschnote sind deutlich enger als die Beziehungen zur Benotung im Fach Mathematik. Von diesem Befundmuster weicht allein die Zuhörkompetenz ab, diese ist in vergleichbarer Stärke mit der Deutsch- wie auch mit der Mathematiknote assoziiert. Gleichzeitig ist die Korrelation zwischen der Deutschnote und der Zuhörkompetenz deutlich geringer als die Zusammenhänge der Deutschnote mit den anderen Kompetenzbereichen (Bremerich-Vos, Böhme & Robitzsch, 2009, S. 214). Dieses Ergebnis kann so interpretiert werden, dass die für diese Domäne charakteristischen Fähigkeitsaspekte im Deutschunterricht einen geringeren Stellenwert haben als die anderen Kompetenzbereiche und sich die Testleistung im Bereich Zuhören in fast identischem Umfang auch in der Mathematiknote widerspiegelt. Es kann also vermutet werden, dass sich das Zuhören in den Noten für beide Fächer als allgemeine Voraussetzung des schulischen Lernens niederschlägt und nicht als spezifischer Gegenstand des Deutschunterrichts behandelt wird. Vor dem Hintergrund der vorschulischen Entwicklung

sprachlicher Fähigkeiten (vgl. Abschnitt 2.6) und dem Kernauftrag der schulischen Vermittlung von Lesefertigkeit und -fähigkeit zu Beginn der Beschulung in der Primarstufe ist dieser Befund durchaus plausibel. Er belegt eindrucksvoll, dass die schulische Förderung der Zuhörkompetenz, wie sie in den Bildungsstandards für das Fach Deutsch im Primarbereich angelegt ist (KMK, 2005a, S. 9f.), noch nicht der gegenwärtigen unterrichtlichen Praxis entspricht und verglichen mit den anderen Kompetenzbereichen ein deutlich geringeres Maß an curricularer Validität besitzt. Auch innerhalb der deutschdidaktischen Diskussion stellt die Beschäftigung mit dem Zuhören einen relativ jungen Gegenstand dar. Aufgrund dieser besonderen Stellung soll die Zuhörkompetenz im ersten empirischen Beitrag dieser Arbeit eine vertiefte Würdigung erfahren.

3.2 Kompetenzbereich I – Sprechen und Zuhören

In den Bildungsstandards der Kultusministerkonferenz im Fach Deutsch für den Primarbereich (KMK, 2005a) umfasst der Kompetenzbereich I *Sprechen und Zuhören* sowohl die produktive als auch die rezeptive mündliche Sprachkompetenz der Schülerinnen und Schüler, wobei der Schwerpunkt der schulischen Förderung dem Sprechen gilt. Ferner wird das Szenische Spielen diesem ersten Kompetenzbereich zugeordnet.

Vereinfachend lassen sich die in den Bildungsstandards genannten Bereiche mündlicher Sprache wie folgt gruppieren:

- Sprechen (Gespräche führen, zu anderen sprechen, über Lernen sprechen),
- Zuhören (verstehend zuhören, Gespräche führen) und
- Szenisch spielen (vgl. KMK, 2005a, S. 9f.).

Aufgrund des interaktiven Charakters mündlicher Sprachverwendung stellt die hierbei vorgenommene Trennung eine starke Vereinfachung der realen Kommunikationssituation dar, da beispielsweise die in den Standards thematisierte Gesprächsführung stets sowohl Aspekte des Sprechens als auch des Zuhörens vereint.

Im Folgenden soll zunächst eine theoretische Einordnung beider mündlicher Sprachkompetenzen erfolgen, da zahlreiche Aspekte sowohl für die mündliche Produktion als auch die mündliche Rezeption von Bedeutung sind. Nachfolgend werden in getrennten Unterkapiteln noch einmal spezifische Gesichtspunkte für das Sprechen und das Zuhören thematisiert.

3.2.1 Aspekte mündlicher Sprachkompetenz

Unmissverständlich legen die Standards den Fokus der Kompetenzbeschreibungen auf die in Schule und Alltag relevante *Kommunikationsfunktion* mündlicher Sprache. Dabei lassen sich insbesondere zwei basale Zielstellungen der Kommunikation extrahieren: Dies ist zum einen der Transport von Information und zum anderen die soziale Interaktion.

So legen die Standards einen großen Wert auf die Gewinnung (Zuhören) und die Vermittlung (Sprechen) von Informationen mit Hilfe mündlicher Sprache. Auch die explizit erwähnten mündlichen Unterrichtsbeiträge der Schülerinnen und Schüler lassen sich dem Transport von Information zuordnen (vgl. KMK, 2005a, S. 8).

Auf der anderen Seite betonen die Bildungsstandards, dass die mündlichen Sprachkompetenzen so ausgebildet werden sollen, dass sie eine gelingende soziale Interaktion gestatten. So heißt es: „Sprechen ist immer auch soziales Handeln“ (KMK, 2005a, S. 8). Letzteres betrifft nicht nur die Mitteilung von auf die eigene Person bezogenen Inhalten in der sozialen Interaktion, also beispielsweise eigene Gedanken und Gefühle, sondern auch eine adäquate Kommunikationsweise als Komponente eines angemessenen Sozialverhaltens. Es soll den Schülerinnen und Schülern ermöglicht werden, Gespräche situations- und adressatengerecht zu führen. Somit wird der für eine adäquate soziale Interaktion relevante Kommunikationsaspekt berücksichtigt, in dem die Bildungsstandards als Ziel formulieren, den Schülerinnen und Schülern eine demokratische Gesprächskultur und eine konstruktive Form der Gesprächsführung zu vermitteln.

Die Bildungsstandards gehen somit davon aus, dass sowohl die produktive als auch die rezeptive mündliche Sprachkompetenz (fast) immer in einer Situation sozialer Interaktion beziehungsweise Kommunikation zum Tragen kommen. Dieser Umstand zeichnet die mündlichen Sprachkompetenzen aus und markiert einen wesentlichen Unterschied zu den schriftsprachlichen Kompetenzen. Man könnte einwenden: Auch wenn man schreibt, schreibt man mit einer bestimmten Absicht und für einen bestimmten Adressaten. Der Produktionsprozess des Schreibens selbst aber hat – anders als das Sprechen – keinen interaktiven Charakter. Ähnlich verhält es sich mit der Rezeption von Schriftsprache, dem Lesen. Auch dieses findet zumeist mit einer bestimmten – mitunter sozial motivierten – Intention statt, in aller Regel liest man aber für sich allein. Das Zuhören jedoch erfolgt einen Großteil der Zeit in einer interaktiven Kommunikationssituation, beispielsweise während man ein persönliches oder ein Telefongespräch führt.

Auch wenn an dieser Stelle nicht behauptet wird, dass die mündlichen Sprachkompetenzen ausschließlich in der interaktiven Kommunikation relevant sind, so sind in

den Bildungsstandards im Zusammenhang mit Sprechen und Zuhören doch *in erster Linie* Aspekte der sozialen Interaktion und somit wesentliche Bestandteile der sozialen Kompetenz angesprochen. Dies ist ein Bruch mit der Forderung der Standards an sich selbst, fachspezifische Kernkompetenzen zu benennen (vgl. Köller, 2008, 2010). So spielt beispielsweise eine konstruktive Gesprächsführung und das Äußern eigener Gedanken auch in anderen Fächern wie Ethik oder Sozialkunde eine bedeutsame Rolle. Insofern kann hinterfragt werden, ob es sich hierbei um eine Kompetenz handelt, die spezifisch für den Fachunterricht Deutsch ist, oder ob die mündlichen Sprachkompetenzen nicht vielmehr eine fächerübergreifende Position einnehmen, da mündliche Unterrichtsbeiträge und das Aufnehmen mündlich vermittelter Information für alle Schulfächer unerlässlich sind.

Eine andere mögliche Sichtweise auf den Bereich „Sprechen" bietet Ehlich (2009) an, indem er schreibt: „In den anderen ‚Fächern' ist das Sprechen Mittel für die Erreichung von deren didaktischen Zielen. Im Deutschunterricht ist die Aneignung dieses Mittels selbst ein zentrales didaktisches Ziel" (S. 11). Somit betrachtet Ehlich (2009) das Sprechen als Unterrichtsgegenstand des Deutschunterrichts, für den übrigen Fachunterricht diene es hingegen nur als Werkzeug. Ganz ähnlich argumentiert Nauwerck (2009), wenn sie schreibt: „Während die Sprache im Deutschunterricht Unterrichtsgegenstand und Medium zugleich ist, werden in allen Fächern Inhalte meist sprachlich transportiert" (S. 260). Eine Verortung des Sprechens in den Standards für das Fach Deutsch ist also durch die Sichtweise gerechtfertigt, dass Sprache und insbesondere das Sprechen und Zuhören im schulischen Kontext stets Medium sind, im Deutschunterricht aber darüber hinaus als inhaltlich relevanter Unterrichtsgegenstand thematisiert und gefördert werden.

Nichtsdestotrotz beinhalten die Kompetenzbeschreibungen des Bereichs *Sprechen und Zuhören* zahlreiche Aspekte, die nicht exklusiv den Deutschunterricht betreffen und teils sogar als Aspekte des allgemeinen Sozialverhaltens gelten müssen, welches weit über den schulischen Kontext hinaus Relevanz besitzt. Inwieweit nun die Vermittlung eines angemessenen und gesellschaftlich akzeptierten sozialen Umgangs Aufgabe der Schule oder sogar speziell des Deutschunterrichts ist, soll an dieser Stelle nicht weiter diskutiert werden.

Es kann festgehalten werden, dass die Bereiche *Sprechen und Zuhören* im schulischen Kontext einen herausragenden Stellenwert einnehmen, da sie zentral für die unterrichtliche Kommunikation und damit den Transport von Wissen in allen Fächern sind. Sie stellen bereits zu Beginn der schulischen Wissensvermittlung eine zentrale Lernvoraussetzung dar. Dieser Umstand mag nicht allen Lehrkräften in seiner ganzen Tragweite bewusst sein, weshalb die Förderung von Sprechen und Zuhören im Deutschunterricht bislang eine deutlich geringere Aufmerksamkeit erfahren hat als die Vermittlung der schriftsprachlichen Fähigkeiten, die als

originärer Auftrag der Beschulung im Primarbereich wahrgenommen werden (vgl. hierzu Büker & Vorst, 2010). Natürlich besteht zwischen mündlichen und schriftlichen Sprachkompetenzen ein wesentlicher Unterschied hinsichtlich ihres Erwerbsstatus zum Zeitpunkt des Schuleintritts (vgl. Abschnitt 2.6). Obwohl die mündlichen Fähigkeiten zum Schuleintritt bei allen Kindern bereits zu einem gewissen Maß ausgebildet sind, kann nicht davon ausgegangen werden, dass alle Grundschulkinder über ein einheitliches und hinreichendes Kompetenzniveau hinsichtlich dieser Fähigkeiten verfügen. Zum einen sollte hier an Kinder nicht-deutscher Herkunftssprache gedacht werden, die in ihrem familiären Kontext möglicherweise noch nicht genügend außerinstitutionelle Lerngelegenheiten für den Aufbau einer soliden Fähigkeitsbasis im deutschsprachigen Sprechen und Zuhören erfahren haben (vgl. Oomen-Welke, 2008; Reich, 2003). Zum anderen ist zu bedenken, dass schulisches Sprechen eine deutlich andere Qualität besitzt als Sprechen im Alltag. Neben dem höheren Stellenwert der Befolgung von Kommunikations- und Interaktionsregeln, die beispielsweise die Vergabe des Rederechts regulieren, kommt hinzu, dass ein zentrales Bildungsziel des Deutschunterrichts der Wechsel von der mündlichen in die schriftliche Modalität und umgekehrt ist (vgl. Ehlich, 2009, S. 9). Mit Bezug auf die Qualität schulischen Sprechens bemerkt Ehlich (2009) ferner: „Sprechen in der Transformation zur Schriftlichkeit bedeutet nicht nur den Wechsel von Diskurs und Text, es bedeutet auch, dass die Schüler/innen sich neben den Textwelten neue Diskurswelten erschließen müssen“ (S. 9). Ebenso betont auch Nauwerck (2009) hinsichtlich schulischen Sprechens:

> Erschwerend kommt dabei das im institutionalisierten Kontext verwendete schulspezifische Register hinzu. Dieses ist oft an der schriftsprachlichen Norm orientiert und unterscheidet sich auf lexikalischer, syntaktischer und pragmatischer Ebene deutlich von der Alltagssprache, so dass es selbst für sprachlich altersangemessen entwickelte Kinder mitunter durchaus eine Herausforderung darstellen kann. (Nauwerck, 2009, S. 260)

Folglich sollte neben der Vermittlung schriftsprachlicher Fähigkeiten auch die weiterführende Förderung mündlicher Fähigkeiten im Deutschunterricht der Primarstufe ihren Platz haben.

Stellt man die hier thematisierten mündlichen und die nachfolgend relevanten schriftlichen Sprachfähigkeiten einander gegenüber, zeigt sich, dass sich auch die Deutschdidaktik über viele Jahre hinweg in erster Linie als eine Didaktik der schriftlichen Sprachkompetenzen verstanden hat, weshalb die systematische Forschung zu den Bereichen Lesen, Textproduktion und Orthografie deutlich weiter voran geschritten ist als die empirische Untersuchung von Sprechen und Zuhören. So konstatieren beispielsweise Behrens und Eriksson (2009): „Die Didaktik der Mündlichkeit steckt noch in den Kinderschuhen“ (S.47). Weiter schreiben die

Autorinnen: Es „wird allenthalben mit Recht ein beträchtlicher Mangel an empirisch gestütztem Wissen sowohl zur individuellen Entwicklung von Sprech-, Zuhör- und Gesprächskompetenz als auch zu unterscheidbaren Komponenten der verschiedenen Bereiche und ihren Zusammenhängen untereinander beklagt“ (Behrens & Eriksson, 2009, S. 73). Bei einer insgesamt unzureichenden Forschungstätigkeit bezüglich der mündlichen Sprachfähigkeiten in der Muttersprache ist der Wissensstand hinsichtlich des Zuhörens noch deutlich defizitärer als im Hinblick auf das Sprechen. Während der frühkindliche Spracherwerb entwicklungspsychologisch sehr gut untersucht und bis zum Grundschulalter detailliert dokumentiert ist (Grimm & Wilde, 1998), beziehen sich Studien und Erläuterungen zum Zuhören vorrangig auf die Zeit vor und kurz nach der Geburt. Auch in der Gestaltung des schulischen Deutschunterrichts nimmt der Kompetenzbereich Zuhören nur einen sehr geringen Stellenwert ein (vgl. Bremerich-Vos, Böhme & Robitzsch, 2009). Entsprechend äußert Bremerich-Vos (2009): „Im Bereich ‚Sprechen und Zuhören’ gibt es, soweit die ‚Muttersprache’ betroffen ist, bislang allenfalls rudimentäre Modelle von Teilfähigkeiten und Fähigkeitsniveaus. Vor allem im Teilbereich ‚Zuhören’ kann auch kaum von einer tradierten Aufgabenkultur gesprochen werden“ (S. 31).

Dass die mündlichen Sprachkompetenzen bislang nicht in zufriedenstellendem Umfang und in hinreichender Qualität untersucht wurden, begründet sich allerdings nicht nur in einem Mangel an didaktischem Interesse. Vielmehr müssen sowohl das Sprechen als auch das Zuhören als Konstrukte charakterisiert werden, die beispielsweise im Rahmen von Large-Scale-Assessments nur schwer operationalisierbar sind. Dies ist – neben technischen Beschränkungen – in erster Linie dem interaktiven Grundcharakter der mündlichen Sprachkompetenzen geschuldet. Sprechen und, in geringerem Maße, auch Zuhören finden – wie oben ausgeführt – primär in Kommunikations- oder Gesprächssituationen statt. Zuhör- und Sprechaufgaben zu konstruieren, die jeglicher Interaktion entbehren und dennoch eine relevante und valide Anforderung darstellen, ist fast unmöglich. Ferner ist die Leistungsbewertung für authentische Aufgaben im Bereich des Mündlichen deutlich problematischer als für das Gebiet der Schriftsprache. Aufgrund des interaktiven Charakters von Sprechen und Zuhören wird der eigene Beitrag beziehungsweise die eigene erbrachte Leistung stets wesentlich durch den Beitrag beziehungsweise die Leistung des Interaktionspartners bestimmt. Dieser Umstand maximiert Probleme der fairen Leistungsbewertung, die durch die Flüchtigkeit mündlicher Sprache ohnehin schwierig ist. Somit wird von Seiten der Deutschdidaktik berechtigterweise die Frage formuliert, ob mündliche Kompetenzen überhaupt im Rahmen standardisierter Leistungsermittlung quantitativ erfassbar sind (vgl. Vogt, 2009).

Inwieweit dies möglich ist, soll in den nachfolgenden Abschnitten thematisiert werden. Alternativ zu einer standardisierten Überprüfung mündlicher Kompetenzen im Rahmen großer

Schulleistungsstudien können für den mündlichen Bereich aber auch kleinere Studien, mit qualitativem oder experimentellem Charakter, zielführend sein, da diese den Besonderheiten mündlicher Sprachkompetenzen eher gerecht werden können.

Wie oben erwähnt umfasst der Kompetenzbereich I in den Bildungsstandards für das Fach Deutsch im Primarbereich (KMK, 2005a) neben den beiden Bereichen des Sprechens und Zuhörens auch Standards, die sich auf den Bereich des Szenischen Spielens beziehen. Dieser Kompetenzbereich ist in hohem Maße durch gestaltendes Handeln und zwischenmenschliche Interaktion gekennzeichnet, weshalb eine Überführung in Testaufgaben für ein Large-Scale Assessment nicht möglich ist. Dies mindert jedoch nicht die Bedeutung, die dieser Bereich für die unterrichtliche Praxis als Gegenstand des Deutschunterrichts hat. Der Vollständigkeit halber soll daher vor der vertieften Auseinandersetzung mit dem Sprechen sowie dem Zuhören das Szenische Spielen in einem kurzen Exkurs thematisiert werden.

3.2.2 Exkurs: Szenisch Spielen

Belgrad und Kollegen (2008) unterscheiden zwei Grundformen des Szenischen Spielens: Zum einen das Rollenspiel, welches in erster Linie auf das Verstehen von Personenkonstellationen und Interaktionsstrukturen abzielt und zum anderen das Darstellende Spiel, für das eine ästhetische Ausgestaltung einer Textvorlage charakterisierend ist. Für beide Spielformen benennen die Autoren so genannte Teilkompetenzen. Hierbei handelt es sich um Inszenierungs- und dramaturgische Kompetenzen (Inszenierungsideen, Raumdramaturgie), sprecherische Kompetenzen (präzise Artikulation, Lautstärke, Stimmführung), körperliche Kompetenzen (u. a. Körperhaltung, Gestik, Mimik) sowie mediale Kompetenzen (Auswahl und Gestaltung von Requisiten, Kostümen, Licht, Ton usw.) (vgl. Belgrad, Eriksson, Pabst-Weinschenk & Vogt, 2008, S. 34 ff.).

Hiervon abweichend gebrauchen Behrens und Eriksson (2009) das Konzept des Szenischen Spielens in einem deutlich weiteren Sinn und fassen verschiedenste Formen mündlicher Kommunikation unter diesem Begriff zusammen. Hierbei stellen sie produktiven Formen des Szenischen Spielens (szenische Improvisation, Rollenspiel), bei denen eine Textvorlage in der Situation spontan generiert wird, reproduktive Formen (Vorlesen, Rezitieren, Darstellendes Spiel), die von einer bestehenden Textvorlage ausgehen, gegenüber (vgl. Behrens & Eriksson, 2009, S. 53). Neben der sprachlichen und stimmlichen Gestaltung sind hier auch Aspekte der nonverbalen Kommunikation und ferner der medialen Präsentation von Bedeutung.

Hinsichtlich der Einschätzung der im Bereich des Szenischen Spielens vorliegenden Kompetenzstände merken Belgrad und Kollegen (2008) an, dass es hier keine eindeutigen Lösungen und Bewertungen im Sinne von „richtig" oder „falsch" geben könne, da es sich um den Bereich der ästhetischen Produktion handele und somit lediglich Kriterien wie Angemessenheit oder Textverträglichkeit zum Tragen kommen könnten. Wesentlich sei es, dass die gewählten Lösungen argumentativ gerechtfertigt werden könnten (vgl. Belgrad, Eriksson, Pabst-Weinschenk & Vogt, 2008, S. 41 f.). So geben die Autoren für die Formulierung von Testaufgaben unter anderem folgendes Beispiel: „Antworten Sie argumentierend auf vier Feedbacks zu Ihrer Sprechgestaltung der Figuren" (vgl. Belgrad et al., 2008, S. 42). Problematisch erscheint aus messtheoretischer Sicht, dass somit die Schlüssigkeit und Überzeugungskraft der mündlichen Argumentation entscheidend für die Bewertung ist. Gemessen wird in diesem Beispiel, wie gut die Schülerin beziehungsweise der Schüler mündlich oder schriftlich argumentieren kann, seine Kompetenz im Bereich des Szenischen Spiels wäre aber nur – der beliebig austauschbare – Gegenstand, an dem die Argumentationskompetenz gezeigt wird.

Der Teilbereich des Szenischen Spielens stellt ein diagnostisch noch nicht erschlossenes Konstrukt dar, welches sich einer standardisierten Leistungsmessung zurzeit noch entzieht. Möchte man potentielle Gegenstände künftiger Forschung an der Schnittstelle zwischen Deutschdidaktik und pädagogisch-psychologischer Diagnostik im Rahmen der empirischen Bildungsforschung benennen, so kann die Konstruktbeschreibung und Diagnostik für das Szenische Spielen als Desideratum vermerkt werden.

3.2.3 Teilbereich Sprechen

Hinsichtlich des Stellenwerts des Kompetenzbereichs Sprechen schreibt Ehlich (2009):

> Die Befähigung zum Sprechen ist eine Grundbefähigung für mündige Bürger, die die Kinder in unserer Gesellschaft werden sollen und für die sie durch die Institution Schule die grundlegende Qualifizierung erhalten. Die entsprechenden Lernzielbestimmungen – und selbstverständlich auch die Testung des Erreichens solcher Lernziele – haben also für unser Gemeinwesen eine basale Qualität. (Ehlich, 2009, S. 10)

Da somit sowohl aus deutschdidaktischer wie auch aus diagnostischer Perspektive das Bedürfnis besteht, das Niveau der Sprechkompetenz der Schülerinnen und Schüler verlässlich bestimmen zu können, soll an dieser Stelle der Versuch unternommen werden, wesentliche Elemente des *Konstrukts* der Sprechkompetenz zu beschreiben und auf zentrale *Herausforderungen bei der Testung* in diesem Bereich hinzuweisen. Aufgrund der Komplexität des Gegenstands kann dieser Versuch

über einen Problemaufriss allerdings nicht hinausgehen und keine erschöpfenden Lösungen anbieten.

Wie auch für andere sprachliche Konstrukte kann hinsichtlich der Testentwicklung für die Erfassung mündlicher Sprachkompetenzen auf einen breiten Erfahrungsschatz aus dem englischen Sprachraum zurückgegriffen werden, wo die Sprachtestung insgesamt und speziell in diesem Bereich auf einem hoch professionellen Niveau stattfindet.

Luoma (2004, S. 5f.) erläutert den Kreislauf der auf das Sprechen bezogenen Testentwicklung und benennt die folgenden Eckpunkte: Bedarfsfeststellung, Entwicklung von Aufgaben und Bewertungskriterien, Testadministration, Bewertung der erbrachten Leistungen sowie Nutzung der ermittelten Kompetenzstände entsprechend des zu Beginn festgestellten Bedarfs. Diese Schritte unterscheiden sich nur wenig von den stets üblichen Etappen der Testentwicklung.

Wie für andere Leistungsdomänen auch ist der Ausgangspunkt der Testung ein hinsichtlich einer Kompetenzbewertung festgestellter Bedarf. Dieser Bedarf muss in Bezug auf Zweck und Art der benötigten Kompetenzeinschätzung möglichst präzise konkretisiert werden. Es folgt eine Phase der Planung und Entwicklung der diagnostischen Aufgaben, in der zunächst definiert werden muss, welches Konstrukt konkret in Testaufgaben überführt werden soll. Die entsprechend entwickelten Aufgaben müssen erprobt und gemäß den empirischen Befunden überarbeitet werden. Spezifisch für die Kompetenzmessung im Bereich Sprechen ist nun, dass neben den Aufgaben auch die Bewertungskriterien und der Testablauf intensiv erprobt und gegebenenfalls revidiert werden müssen. Eine detailliert geplante und genau organisierte Testadministration ist für den Erfolg einer Testung der Sprechkompetenz von ausschlaggebender Bedeutung und nimmt damit einen deutlich prominenteren Stellenwert als beispielsweise bei der Überprüfung der Lesekompetenz ein. In der eigentlichen Testsituation treten die Testteilnehmer entweder miteinander oder mit einem Testleiter in Interaktion und liefern hierbei Proben ihrer Sprechkompetenz. Die Überprüfung der Sprechkompetenz erfordert also einen *interaktiven Prozess,* der bei der Testung keiner anderen sprachlichen Kompetenz in diesem Ausmaß in Erscheinung tritt. Die in den Interaktionssituationen generierten Sprechproben werden beispielsweise als Audio- oder Videoaufzeichnungen dokumentiert und ggf. transkribiert, da die zu beurteilende Leistung flüchtig ist und für eine zeitlich nachgelagerte Bewertung beziehungsweise die Überprüfung einer bereits erfolgten Bewertung verstetigt werden muss. Anschließend wenden Urteiler die erprobten und gegebenenfalls revidierten Beurteilungskriterien auf die aufgezeichneten Sprechproben an. Erfolgt die Beurteilung bereits während der Testsituation, ist es erforderlich, zusätzlich zum Testleiter eine weitere Person einzubeziehen, da

es kaum gelingt, gleichzeitig den Testablauf zu steuern, als Interaktionspartner aufzutreten und zusätzlich eine objektive und unverfälschte Einschätzung der Sprechleistung des Interaktionspartners abzuliefern. Die resultierende Kompetenzeinschätzung kann dann entsprechend des eingangs festgestellten Bedarfs dokumentiert und in einer geeigneten Form der Rückmeldung kommuniziert werden.

In dem für die vorliegende Arbeit relevanten Kontext besteht der diagnostische Bedarf in einem verlässlichen Instrument, welches eine Einschätzung gestattet, ob und inwieweit die in den Bildungsstandards beschriebenen Sprechkompetenzaspekte bei den Schülerinnen und Schülern des Primarbereichs vorliegen. Wie die resultierende Kompetenzeinschätzung genau beschaffen sein sollte – ob beispielsweise ein einzelner Testwert einem Kompetenzprofil vorzuziehen ist – und welchem konkreten Zweck die Kompetenzfeststellung dienen soll, – etwa der Unterrichtsentwicklung, einer optimierten Lehrerausbildung oder ausschließlich dem Bildungsmonitoring – ist bislang nicht hinreichend geklärt. Da die Zweckbestimmung und die Beschaffenheit des Testinstruments eng miteinander verbunden sind, sollte daher zunächst in Abstimmung von Bildungsadministration, Psychometrie und Deutschdidaktik die diagnostische Zielstellung konkretisiert werden.

Bei der standardisierten Leistungsprüfung im Bereich Sprechen müssen zahlreiche Herausforderungen bewältigt werden. Diese erstrecken sich vom angestrebten interaktiven Charakter der Testsituation und der Entwicklung entsprechender Aufgaben über die Flüchtigkeit der Testleistung hin zur reliablen Einschätzung der Sprechproben durch geschulte Beobachter, bei der sich neben dem Wunsch nach einer möglichst hohen intra- und intersubjektiven Übereinstimmung der Urteiler auch die Frage nach geeigneten Kriterien der Leistungsbeurteilung stellt.

Die Frage nach zweckdienlichen Kriterien für die Beurteilung von Sprechleistungen spiegelt die Ansicht wider, dass mündliche Kommunikation nur eingeschränkt als „richtig“ oder „falsch“ klassifiziert werden kann. Daher ist die Frage der jeweiligen situativen Angemessenheit zu berücksichtigen (vgl. Belgrad et al., 2008, S. 31).

Dennoch ist es möglich, aus einer Konstruktbeschreibung wesentliche Aspekte abzuleiten, die hinsichtlich des Grades ihrer Bewältigung beurteilt werden können und auf diese Weise zu einer objektiven und reliablen Einschätzung der Sprechkompetenz führen.

Luoma (2004, S. 9ff.) beschreibt Sprechen aus der Perspektive der angewandten Linguistik und benennt unter anderem die folgenden Aspekte als zentrale Eckpfeiler der Konstruktbeschreibung:

Aussprache und Sprachklang: Die Aussprache beziehungsweise der Sprachklang umfassen die bewusste und gezielte Variation von Lautstärke, Sprechgeschwindigkeit, Tonhöhe,

Sprechpausen, Betonung und Intonation. Bei der Beurteilung dieser Kriterien muss jeweils entschieden werden, was als angestrebter Standard gelten soll, da in jeder Sprache mehrere „Normvarianten" parallel existieren können. Sinnvoller erscheint daher das Kriterium, eine Kommunikationssituation effektiv gestalten zu können und beim Sprechen *verständlich* zu sein. Mitunter wird in diesem Zusammenhang auch ausschließlich die Korrektheit der Aussprache bewertet. Verständlichkeit umfasst aber wesentlich mehr als eine korrekte Aussprache, beispielsweise die bereits angesprochene gezielte Variation von Sprechgeschwindigkeit, Intonation, Betonung und Sprechrhythmus. Der Aspekt des Sprachklangs ist für Muttersprachler deutlich weniger relevant als für Kinder nicht-deutscher Herkunftssprache.

Grammatikalität: Der auf das Sprechen bezogene Kompetenzstand deutschsprachiger Schülerinnen und Schüler, ebenso wie der von Kindern nicht-deutscher Herkunftssprache, kann anhand der Vielfalt und der Korrektheit grammatikalischer Konstruktionen in der gesprochenen Sprache beurteilt werden (Luoma, 2004, S. 11). Der jeweilige Leistungsstand kann anhand dieses Kriteriums sinnvoll eingeschätzt werden, da sich grammatikalische Konstruktionen mit zunehmender Kompetenz von der Wiederholung weniger, einfacher Konstruktionen zu vielfältigen und komplexeren Strukturen entwickeln. Grammatikalisch hoch komplexes Sprechen kann allerdings eine Beeinträchtigung der Verständlichkeit mit sich bringen, weshalb nicht davon auszugehen ist, dass höchste grammatikalische Komplexität Ausweis höchster Kompetenz ist. Neben dem Grad der Komplexität sollten auch grammatikalische Fehler Berücksichtigung finden, da zumeist mit steigender Kompetenz eine deutliche Abnahme der Fehlerzahl und -dichte festzustellen ist.

Von wesentlicher Bedeutung bei der Beurteilung der Grammatikalität während des Sprechens ist die Tatsache, dass nicht die Maßstäbe schriftlicher Grammatikalität an mündliche Äußerungen angelegt werden dürfen. Ferner ist relevant, inwieweit es sich bei dem zu beurteilenden Sprechbeitrag um vorbereitetes oder unvorbereitetes Sprechen handelt. Mit zunehmender Vorbereitung ist auch hinsichtlich der Grammatikalität höhere Komplexität und eine geringere Fehlerzahl zu erwarten.

Wortschatz: In Abhängigkeit von der konkreten Kommunikationssituation kann sowohl eine einfache, begrenzte als auch eine vielfältige und komplexe Wortwahl angemessen und somit ein Ausweis eines hohen Kompetenzniveaus sein. Ferner ist der Gebrauch von Abtönungspartikeln und Füllwörtern nicht unbedingt ein Ausdruck mangelnder Kompetenz, da die wahrgenommene Sprechflüssigkeit beispielsweise durch die Verwendung von Verzögerungsphänomen, Abtönungs- und Diskurspartikeln positiv beeinflusst wird (Hasselgren, 2002).

Fehler und Versprecher: Normales Sprechen beinhaltet eine Vielzahl an kleineren Fehlern und Versprechern, die auch hochkompetenten Sprechern unterlaufen können. Daher müssen diese „normalen" Fehler von solchen Fehlern abgegrenzt werden, die grammatikalische Phänomene betreffen und Ausdruck mangelnder Sprachkompetenz im Deutschen sind.

Diese möglichen Kriterien der Einschätzung von Sprechbeiträgen stellen nur eine Auswahl von einigen wenigen unter zahlreichen denkbaren Eigenschaften dar. Sie dienen der Veranschaulichung, dass es durchaus möglich ist, auch für die mündliche Sprachproduktion bewertbare Kriterien zu definieren.

Diese Überlegung trifft insbesondere auf jene Standards zu, die Sprechen nicht primär als Bestandteil sozialer Interaktion thematisieren, sondern beispielsweise darauf abzielen, Lernergebnisse zu präsentieren (KMK, 2005a, S.10). Bereits in der Grundschule sollen die Schülerinnen und Schüler möglichst vielfältig Gelegenheit erhalten, Referate und Präsentationen zu halten und hierbei zu üben, vor und zu anderen zu sprechen (vgl. Belgrad et al., 2008; Pabst-Weinschenk, 2005). Auch die Überprüfung und Bewertung des Sprechens in Präsentationssituationen muss sich an dem Grad der Angemessenheit orientieren. Dieser wiederum wird maßgeblich durch den jeweiligen Kontext bestimmt (vgl. Belgrad et al., 2008, S. 31).

Nichtsdestotrotz ist Sprechen in aller Regel Bestandteil einer Kommunikationssituation und Ausdruck sozialer Interaktion. Selbst wenn das Sprechen aus testdiagnostischer Perspektive gelegentlich so beschrieben wird, als würde es sich dabei um die beurteilbare Leistung jeweils eines Individuums handeln, steht außer Frage, dass alltägliches und schulisches Sprechen primär in sozialen Interaktionen stattfindet. Hierbei fungiert jeder Teilnehmer abwechselnd als Sprecher und Hörer und die entstehende Gesprächssequenz ist das Produkt der Beiträge aller Teilnehmer. Dies testdiagnostisch adäquat umzusetzen ist und bleibt die zentrale Herausforderung für die Operationalisierung des Kompetenzkonstrukts Sprechen.

3.2.4 Teilbereich Zuhören

Gemessen an der Relevanz der verschiedenen Sprachkompetenzen für das schulische Lernen während der Grundschulzeit nimmt die Zuhörkompetenz den wahrscheinlich größten Stellenwert ein. Belgrad und Kollegen (2008) konstatieren, dass „Schüler(innen) bis zu zwei Drittel der Unterrichtszeit zuhören müssen" (Belgrad et al., 2008, S. 20). Dieser Umstand verdeutlicht, dass ein Großteil des Lerninputs als mündliche Informationsvermittlung erfolgt, weshalb Schülerinnen und Schüler in die Lage versetzt werden müssen, ihr Zuhörverhalten selbständig und kompetent zu steuern.

Zuhören zu können umfasst dabei mehr als eine gelingende akustische Wahrnehmung von Schallwellen. Im schulischen Kontext bedeutet Zuhörkompetenz auch die *Bereitschaft* zuzuhören, die kognitive Verarbeitung des sprachlichen Inputs und die Reflexion von Sprecher- und Situationsmerkmalen vor dem Hintergrund gesellschaftlicher und persönlicher Erwartungen (vgl. Behrens & Eriksson, 2009).

Inwieweit für die Entwicklung und Bewahrung der Zuhörkompetenz eine explizite schulische Förderung erforderlich ist, scheint zunächst offen. Zwar beklagen Lehrpersonen, dass die Fähigkeit der Schülerinnen und Schüler, aufmerksam und konzentriert zuzuhören ungenügend ausgeprägt ist und innerhalb der Schülerschaft eine (zu) große Varianz aufweist (Behrens, Böhme & Krelle, 2009). Dennoch spielen deutschdidaktische Überlegungen zum Bereich Zuhören bislang nur eine untergeordnete Rolle (vgl. Krelle, 2010). Momentan findet innerhalb der Deutschdidaktik allerdings ein Wandel statt, der dem Zuhören als Kompetenzbereich einen deutlich höheren Stellenwert einräumt. Entsprechend stellen beispielsweise Imhof und Bernius (2010) fest, „dass in der Zuhörforschung Fortschritte gemacht worden sind“ (S. 13). Neben empirischen Forschungsarbeiten zur Diagnostik der muttersprachlichen Zuhörkompetenz (Behrens, Böhme & Krelle, 2009) finden sich in jüngster Vergangenheit auch fundierte deutschsprachige Untersuchungen zu didaktischen Fragen, die sich nicht nur empirischer Methoden bedienen, sondern darüber hinaus den Blick in benachbarte Disziplinen wagen. Beispielgebend sind hier etwa die Untersuchungen von Gailberger zur Leseförderung bei schwachen Schülerinnen und Schülern mit Hilfe von Hörbüchern (Gailberger, 2008, 2010).

Ähnlich wie für den Bereich der Lesekompetenz – und möglicherweise bedingt durch die enge Verwandtschaft zu dieser anderen rezeptiven Sprachkompetenz – gibt es auch zum Zuhören bereits einige kognitionspsychologische und psycholinguistisch motivierte Bemühungen um die Erarbeitung eines theoretischen Kompetenzmodells. Prominent sind hier insbesondere die Arbeiten von Imhof, die Zuhören als einen mehrstufigen Informationsverarbeitungsprozess versteht, bei dem in den Phasen der Intentionsbildung, der Selektion, der Organisation und der Integration die Selbstregulation des Zuhörenden eine ausschlaggebende Rolle spielt (vgl. Imhof, 2003, 2010).

Trotz dieser fruchtbaren Anstrengungen der aktuellen deutschsprachigen Zuhörforschung, deren Stärken in ihrer Interdisziplinarität und der Integration von empirischen Forschungsmethoden in fachdidaktische Ansätze liegen, bleiben nach wie vor noch zahlreiche Fragen unbeantwortet, beispielsweise die der Entwicklung der Zuhörkompetenz über die Lebensspanne.

In den Bildungsstandards im Fach Deutsch für den Primarbereich wird die Zuhörkompetenz unter den Bereich der Mündlichkeit subsumiert, wobei der Fokus der Bildungsstandards eindeutig auf dem Sprechen liegt. In den allgemeinen Erläuterungen der einzelnen Kompetenzbereiche heißt es zur Zuhörkompetenz der Schülerinnen und Schüler lediglich: „Sie […] hören aufmerksam und genau zu, nehmen die Äußerungen anderer auf und setzen sich mit diesen konstruktiv auseinander“ (KMK, 2005a, S. 8). Die formulierten Substandards beschränken sich auf das verstehende Zuhören, und zwar konkret darauf, Inhalte zuhörend zu verstehen, gezielt nachzufragen und Verstehen und Nicht-Verstehen zum Ausdruck zu bringen (KMK, 2005a, S. 10). Bei genauerer Betrachtung müssen auch die beiden letztgenannten Teilkompetenzen als *produktive* Aspekte verbaler beziehungsweise nonverbaler Kommunikation verstanden werden und sind somit eher nicht Konstruktbestandteil des *rezeptiven* Zuhörens. Gegeben die Tatsache, dass es also beim Zuhören als rezeptiver Sprachkompetenz in den Bildungsstandards ausschließlich darum geht, *Inhalte zuhörend zu verstehen,* sollte in dem Bemühen um möglichst große begriffliche Klarheit nicht von Zuhören, sondern von *Hörverstehen* gesprochen werden. Geht man davon aus, dass Textverstehen ein vom Medium – zumindest partiell – unabhängiges Konstrukt ist, können Forschungsergebnisse aus der Leseverstehensforschung, beispielsweise zu schwierigkeitsgenerierenden Merkmalen von Texten und Aufgaben, hilfreiche Ansatzpunkte für die Untersuchung des Hörverstehens liefern. Diesem Ansatz folge ich mit meinen Kollegen Alexander Robitzsch und Anne-Kathrin Busè im nachfolgenden Beitrag.

I. Beitrag 1: Zur Abgrenzung des Hörverstehens gegenüber dem Leseverstehen mit Hilfe schwierigkeitsbestimmender Merkmale bei der Entwicklung von Testaufgaben

Autoren:
Katrin Böhme / Alexander Robitzsch / Anne-Kathrin Busè

Erschienen in:
V. Bernius & M. Imhof (Hrsg.) (2010), *Zuhörkompetenz in Unterricht und Schule.* (S. 81-104) Göttingen: Vandenhoeck & Ruprecht.

Dieser Beitrag beschäftigt sich mit der Operationalisierung des Hörverstehens für Testungen im Rahmen von Large-Scale-Assessments. Hierbei verbinden wir theoretische Überlegungen zu Konstruktoperationalisierung und schwierigkeitsbestimmenden Merkmalen mit Ergebnissen eigener empirischer Forschung aus dem Primarbereich.

I.1 Theoretische Grundlagen der Konstruktdefinition des Hörverstehens

I.1.1 Aspekte der Konstruktoperationalisierung

Das Hörverstehen ist eine rezeptive Sprachkompetenz, die im Kontext großer Schulleistungsstudien als latentes Kompetenzkonstrukt verstanden wird, das es in Aufgaben zu überführen gilt, um das Maß vorhandener Kompetenz quantifizieren zu können. In der pädagogisch-psychologischen Diagnostik sind Konstrukte gedankliche beziehungsweise theoretische Konstruktionen, deren Ausprägung nicht direkt beobachtbar ist. Daher muss aus anderen, der Beobachtung zugänglichen Sachverhalten (so genannten Indikatoren) auf das Konstrukt zurückgeschlossen werden. Die Übertragung eines Konstrukts in prüfbare Indikatoren bezeichnet man als Operationalisierung. Um eine Operationalisierung beispielsweise in Form von Testaufgaben leisten zu können, ist es zunächst erforderlich, ein genaues Verständnis davon zu entwickeln, welche Aspekte als definierende Bestandteile des Konstrukts gelten sollen, damit sich diese (und nur diese) in den Indikatoren wiederfinden.

In diesem Zusammenhang kann bei der Definition eines Hörverstehenskonstrukts entweder eine kompetenzbasierte oder eine aufgabenbasierte Sichtweise eingenommen werden. Der kompetenzbasierte Ansatz setzt bei den Teilfähigkeiten und Subkompetenzen an, die beherrscht werden müssen, um Hörverstehenskompetenz zeigen zu können. Dagegen geht der aufgabenbasierte Ansatz bei der Bestimmung des Konstrukts von den Hörverstehensaufgaben aus, die von den Schülerinnen und Schülern erfolgreich bearbeitet werden sollen.

Im Rahmen der Operationalisierung der Bildungsstandards für das Fach Deutsch im Primarbereich ist nun problematisch, dass in den Formulierungen der Kultusministerkonferenz (KMK, 2005, S. 10) sowohl die Subkompetenzen oder Teilfertigkeiten, die das Kompetenzkonstrukt „Hörverstehen“ ausmachen, als auch die Vorgabe von Aufgabenbeispielen, die von den Schülerinnen und Schüler gelöst werden sollen, nicht konkretisiert werden. So wird in den Bildungsstandards lediglich der Aspekt „verstehend zuhören“ mit der Subkompetenz

„Inhalte zuhörend verstehen" genannt.[1] An diese wenig präzisen Ausführungen schließt sich nun unmittelbar die Frage an, wie die Konkretisierung des Konstrukts „verstehend Zuhören", nachfolgend kurz als „Hörverstehen" bezeichnet, aussehen könnte.

Wählt man den eher theoriegeleiteten Weg einer kompetenzbasierten Konstruktdefinition, so ist zu fragen, welche Teilkompetenzen für das Konstrukt Hörverstehen relevant sind und wie diese gegenüber anderen Konstrukten, wie der Lese- oder der Schreibkompetenz, abgegrenzt werden können.

Buck spricht für eine Konstruktcharakterisierung des Hörverstehens folgende Empfehlungen aus (Buck, 2001, S. 113):

- Das Hauptaugenmerk sollte auf der Erfassung solcher sprachlicher Kompetenzen liegen, die für das Hörverstehen einzigartig sind. Abgeprüft werden sollten also schnelle, automatisierte Verarbeitungsprozesse von Hörtexten, die den Besonderheiten gesprochener Sprache gerecht werden.
- Es ist besonders wichtig, dass getestet wird, ob der Zuhörer basale linguistische Informationen in einer Vielzahl verschiedenartiger Hörtexte zu verschiedenen Themen erfassen kann.
- Es sollten auch längere Texte und Gespräche Verwendung finden, da hier auch Strategien hinsichtlich des Umgangs mit großen Informationsmengen und dauerhaft erforderlicher Aufmerksamkeit zur Anwendung kommen können.
- Es ist wesentlich, nicht nur das Erkennung von wörtlicher Bedeutung zu überprüfen, sondern auch, ob Bedeutungen schlussfolgernd erfasst werden können. Diese erschlossenen Bedeutungen müssen sich aber aus dem Text beziehungsweise linguistischen Hinweisen heraus ergeben.
- Auch vorwissensbasierte Schlussfolgerungen sind wichtig, allerdings muss hier sichergestellt werden, dass alle Testteilnehmer über ein vergleichbares Maß an Vorwissen verfügen. Ist dies nicht gewährleistet, sollten sich diese wissensbasierten Schlussfolgerungen nur auf vorab im Text zur Verfügung gestelltes Wissen beziehen.
- Abschließend sollten alle Aspekte einbezogen werden, die sich auf sprachliches Wissen und Können beziehen. Alle Aspekte, die sich auf allgemeine kognitive Fähigkeiten beziehen, sollten hingegen ausgeschlossen werden.

[1] Die beiden weiteren Subkompetenzen „D-1.3.2 – gezielt nachfragen" und „D-1.3.3 – Verstehen und Nichtverstehen zum Ausdruck bringen" werden im Folgenden nicht berücksichtigt, da sie bislang nicht in Testaufgaben umgesetzt wurden und zum jetzigen Zeitpunkt im Rahmen der aktuellen diagnostischen Möglichkeiten als nicht operationalisierbar gelten müssen.

I.1.2 Muttersprachliches versus fremdsprachliches Hörverstehen

Erschwert wird die Konkretisierung des Hörverstehenskonstrukts auch dadurch, dass sich weite Teile der deutschsprachigen, aber auch der einschlägigen englischsprachigen Literatur nur auf das Hörverstehen in Fremdsprachen und nicht auf jenes in der Muttersprache beziehen. Inwieweit diese beiden Kompetenzen, also das Hörverstehen in der Muttersprache und das Hörverstehen in der Fremdsprache, gleichgesetzt werden können, wird in der Forschung bislang nur angerissen. Auch liegen für den Bereich des Large-Scale-Assessments keine Ergebnisse zum Hörverstehen in der deutschen Sprache vor. Es bleibt somit zusätzlich fraglich, inwieweit Befunde zum englisch-muttersprachlichen Hörverstehen auf das deutsch-muttersprachliche Hörverstehen übertragen werden können, was vorauszusetzen wäre, würde man auf Ergebnisse aus der englischsprachigen Literatur zurückgreifen wollen.

In der englischsprachigen Literatur wird darauf hingewiesen, dass sich fremdsprachliches und muttersprachliches Hörverstehen nicht im eigentlichen Prozess, sondern vorrangig hinsichtlich des begrenzten beziehungsweise unbegrenzten Sprachwissens unterscheiden (Faerch & Kasper, 1986). Wesentlich scheint in diesem Zusammenhang der Automatisierungsgrad der Sprachverarbeitung. Während die Verarbeitung der gesprochenen Sprache in der Muttersprache vollständig automatisch abläuft, ist die Verarbeitung in einer Fremdsprache gar nicht oder in deutlich geringerem Ausmaß automatisiert. Dadurch, dass lexikalische und grammatische Merkmale der gesprochenen Sprache sehr viel Aufmerksamkeit in Anspruch nehmen und bewusst verarbeitet werden müssen, leidet das Hörverstehen in dem Sinne, dass die Erfassung von Bedeutung deutlich eingeschränkt ist (Lynch, 1998). Sprachwissen in der Muttersprache ist also implizit beziehungsweise prozedural und selten deklarativ.

> In our first language, we all know many complex rules and can use them correctly without being aware of them, and usually we could not explain these rules if we tried. The cognitive processes that are involved in first-language understanding are almost entirely automatic. [...] When we hear something in our first language we just automatically understand. (Buck, 2001, S. 49)

In einer Zweit- oder Fremdsprache erreichen wir nur selten ein ähnlich hohes Fähigkeitsniveau wie in der Muttersprache. Dies zeigt sich zum Beispiel bei der kritischen Grenze der Sprechgeschwindigkeit, die problemlos verstanden werden kann und die für Nicht-Muttersprachler geringer als für Muttersprachler ist (Buck, 2001).

Die wenigen verfügbaren empirischen Ergebnisse legen den Schluss nahe, dass das Hörverstehen in der Muttersprache nicht mit dem Hörverstehen in der Fremdsprache gleichgesetzt werden kann und daher differenzierte Konstruktbeschreibungen angezeigt sind.

I.1.3 Konstruktkonfundierung

Bei der Überführung eines Kompetenzkonstrukts in Testaufgaben können nach Messick (1989, 1994) zwei Fehlerarten auftreten: Zum einen können wesentliche Aspekte des theoretisch definierten Konstrukts bei der Operationalisierung nicht im Test repräsentiert sein, diese Operationalisierung wäre dann unvollständig (*construct-underrepresentation*). Zum anderen kann der Test Aspekte messen, die nicht in der theoretischen Konstruktdefinition enthalten sind und daher konstruktirrelevante Varianz darstellen (*construct-irrelevant variance*).

Hinsichtlich der konstruktirrelevanten Varianz ist bei der Operationalisierung des Hörverstehens das Problem der Konstruktkonfundierung mit der Lese- und Schreibkompetenz zu beachten. Die Erfassung einer rezeptiven Kompetenz wie des Zuhörens kann nicht direkt, sondern nur über den Umweg einer produktiven sprachlichen Äußerung erfolgen. Aus testmethodischen Einschränkungen heraus handelt es sich hierbei im Verlauf der Messung im Normalfall um eine schriftsprachliche und keine mündliche Äußerung der Schülerinnen und Schüler. Zur Testung der Zuhörkompetenz wird also eine Antwort provoziert, die bereits an sich eine (andere) zu testende Kompetenz darstellt. Ein Ausbleiben einer Antwort wird auf ein Defizit im Kompetenzbereich „Zuhören" zurückgeführt, ohne dass dies notwendigerweise der Fall sein muss (sondern ebenso gut durch Defizite in der Schreibkompetenz erklärt werden könnte). Ein ganz ähnliches Problem besteht hinsichtlich der rezeptiven Kompetenz des Lesens. Beim Lesen der Instruktionen, der Aufgabenstellungen und den geschlossenen Antwortoptionen wird eine gut ausgebildete Lesekompetenz vorausgesetzt. Ist diese nicht gegeben, scheitert das Kind möglicherweise am Lesen und nicht an der korrekten Bearbeitung einer Aufgabenstellung zum Hörverstehen. Das Ausbleiben der korrekten Antwort würde aber wiederum einem Defizit im Hörverstehen angelastet werden. Eine Möglichkeit, diese Konfundierung zu umgehen, ist es, nicht nur den Stimulustext, sondern auch sämtliche Instruktionen und Aufgabenstellungen akustisch darzubieten, was im Rahmen der Evaluierung der Bildungsstandards für die Sekundarstufe I bereits umgesetzt wird.

I.1.4 Abgrenzung des Hörverstehens gegenüber dem Leseverstehen

Der Fokus empirischer Studien zur Diagnostik (rezeptiver) sprachlicher Kompetenzen liegt auf der Erforschung des Leseverstehens. Hierfür liegen gesicherte empirische Erkenntnisse zu verschiedensten individualdiagnostischen Verfahren sowie auch für das Large- Scale Assessment vor. Die Operationalisierung des (muttersprachlichen) Hörverstehens ist hingegen wenig untersucht. Die naheliegende Fragestellung ist demnach: Sind sich Hör- und Leseverstehen als

rezeptive Sprachkompetenzen so ähnlich, dass eine Unterscheidung beider Konstrukte gar nicht notwendig ist? Ist also die Leseverstehensleistung so eng mit der Hörverstehensleistung assoziiert, dass eine getrennte Operationalisierung keinen zusätzlichen Informationsgewinn liefert? Könnte die Leseverstehenskompetenz somit als Indikator für die Hörverstehenskompetenz fungieren?

In der jüngeren kognitionspsychologischen Forschung wird diskutiert, dass es sich beim Hör- und Leseverstehen weder um völlig identische noch um gänzlich verschiedene Prozesse handelt (vgl. z. B. Kürschner & Schnotz, 2008). Die kognitiven Verarbeitungsprozesse beim Hör- und Leseverstehen können demnach weder als ausschließlich modalitätsunabhängig noch als ausschließlich modalitätsspezifisch gelten. Kürschner und Schnotz (2008) stellten ein integriertes kognitionspsychologisches Modell des Lese- und Hörverstehens vor, welches davon ausgeht, dass menschliche Informationsverarbeitung auf einer Interaktion von Prozessen auf niedrigeren (sensorische Register) und höheren Verarbeitungsebenen (Arbeitsgedächtnis und Langzeitgedächtnis) beruht. Beim Verstehen von auditiv oder visuell dargebotenen Texten werden im Arbeitsgedächtnis mentale Repräsentationen gebildet, die die sprachlichen Informationen eines Textes, dessen semantischen Gehalt sowie eine Integration von semantischem Gehalt und Weltwissen darstellen. Für das Hör- und Leseverstehen wird nun angenommen, dass insbesondere die Wahrnehmung gehörter beziehungsweise gelesener Texte sowie deren Verarbeitung in den sensorischen Registern modalitätsspezifisch sind. Für das Verständnis des semantischen Gehalts und die Integration semantischen Vorwissens aus dem Langzeitgedächtnis wird hingegen die Existenz von sowohl modalitätsspezifischen als auch modalitätsunabhängigen Prozessen angenommen. Kürschner und Schnotz (2008) gehen auch davon aus, dass die modalitätsspezifischen Prozesse auf unteren Verarbeitungsebenen nicht zu qualitativ unterschiedlichen Prozessen auf den höheren Verarbeitungsebenen führen (vgl. auch Leucht, Retelsdorf, Möller & Köller, 2010).

Ein Gleichsetzen der beiden rezeptiven Sprachkompetenzen würde es allerdings unmöglich machen, die Spezifika des Hörverstehens gesprochener Sprache vom Leseverstehen geschriebener Sprache abzugrenzen. Dass diese Besonderheiten existieren, ist in der Literatur jedoch fraglos Konsens (vgl. z. B. Buck, 2001; Grotjahn, 2005). Einige wesentliche Aspekte sind die folgenden:

- Phonologische Modifikationen treten als solche nur bei gesprochener Sprache auf und verändern gesprochene gegenüber geschriebener Sprache.
- Gesprochene Sprache ist redundant. Fehlende oder unklar gesprochene Wörter werden aufgrund der Redundanz der Sprache durch unser Vorwissen ersetzt.

- Sprechgeschwindigkeit und Klarheit der Aussprache werden an das Vorwissen des Gesprächspartners angepasst. Man spricht schneller und undeutlicher, wenn der Gesprächspartner über gleiches Wissen verfügt. Wichtige, nicht redundante Wörter werden langsamer und klarer ausgesprochen.
- Wesentliches Merkmal gesprochener Sprache ist die Vergänglichkeit. Gesprochenes kann nicht (identisch) wiederholt werden, sondern muss bei einmaligem Hören in Echtzeit verarbeitet werden. Dies verdeutlicht die Bedeutsamkeit der Erinnerung an das Gesprochene.
- Aufgrund der hohen Sprechgeschwindigkeit (ca. 3 Wörter pro Sekunde) muss die Verarbeitung automatisch, also ohne die Beteiligung aktiver, bewusster Kontrollprozesse erfolgen.
- Propositionen sind in gesprochener Sprache kürzer und hinsichtlich der verwendeten Syntax weniger komplex. In gesprochener Sprache werden Propositionen durch Konjunktionen (und, oder, aber) verknüpft, in der Schriftsprache erfolgt die Verknüpfung auf komplexere Weise.
- In gesprochener Sprache treten Pausen, Versprecher, Füllwörter und Wiederholungen auf.
- Geschriebene Sprache orientiert sich deutlicher an der Standardsprache, ist formalisierter und ist hinsichtlich der Grammatik korrekter. Gesprochene Sprache ist persönlicher, stärker emotional gefärbt und weniger präzise.
- Wesentlich sind auch Aspekte der Interaktion (Rückfragen, Antwortverhalten etc.), die in dieser Form nur in der gesprochenen Sprache auftreten.

I.2. Schwierigkeitsbestimmende Merkmale in der Entwicklung von Testaufgaben

Für Aufgabenstämme (Lese- bzw. Hörtexte) und Items können verschiedene Merkmale benannt werden, die einen Einfluss auf die Schwierigkeit der Items einer Aufgabe haben. Solche Merkmale werden im Folgenden als schwierigkeitsbestimmende (Aufgaben-)Merkmale bezeichnet. In der Literatur finden sich bereits verschiedene Forschungsarbeiten, die für das Hör- und Leseverstehen schwierigkeitsbestimmende Merkmale definiert und empirisch untersucht haben (vgl. z. B. Buck & Tatsuoka, 1998; Freedle & Kostin, 1993; Grotjahn, 2005; Nold & Rossa, 2007a, 2007b).

Als Beweggründe für die Analyse schwierigkeitsbestimmender Aufgabenmerkmale werden in der Literatur die folgenden benannt: Schwierigkeitsbestimmende Merkmale

- können dazu beitragen, die Varianz in bereits vorliegenden empirischen Daten theoriegeleitet zu erklären (Freedle & Kostin, 1993, S. 166; Buck & Tatsuoka, 1998, S. 120).
- können im Prozess der empirischen Überprüfung erstellter Aufgaben „Unterschiede in den Schwierigkeiten und Trennschärfen der Items erklären helfen und damit den Prozess der Itemrevision erleichtern und zugleich auch zur Konstruktvalidierung beitragen“ (Grotjahn, 2005, S. 124).
- werden auch untersucht, um in der künftigen Aufgabenentwicklung diejenigen Merkmale, die sich als relevant erwiesen haben, systematisch so zu manipulieren, dass zielgerichtet Items einer intendierten Schwierigkeit entstehen (vgl. Grotjahn, 2000, S. 7, 2005, S. 133).
- können dazu dienen, „für jeden Testteilnehmer ein detailliertes, individuelles Fähigkeitsprofil zu ermitteln“ (Grotjahn, 2005, S. 124).
- können zur Definition der Grenzen in Kompetenzstufenmodellen eingesetzt werden (vgl. Hartig, 2007; Nold & Rossa, 2007a, 2007b; Willenberg, 2007).

Um für die hier vorgestellte Studie, die sich auf Hör- und Leseverstehensaufgaben für den Primarbereich stützt, mit einer möglichst erschöpfenden Menge an relevanten Merkmalen arbeiten zu können, wurde zunächst eine Literaturrecherche durchgeführt. Es wurden Forschungsarbeiten identifiziert, die für das Lese- und/oder Hörverstehen Untersuchungen zu schwierigkeitsbestimmenden Merkmalen durchgeführt haben (Buck & Tatsuoka, 1998; Freedle & Kostin, 1993; Grotjahn, 2000, 2005; Köster, 2005; Nold & Rossa, 2007a, 2007b; Willenberg, 2007).

Aus diesen Arbeiten, die zumeist Merkmale auf *Textebene, Itemebene* sowie *Text-Item-Interaktionsebene* untersuchen, konnten wir eine große Anzahl von Merkmalen adaptieren. Die Unterscheidung der verschiedenen Merkmalsebenen haben wir übernommen, wobei eine eindeutige Zuordnung mitunter schwierig scheint. Einen wichtigen Stellenwert unter den gewählten Aufgabenmerkmalen nehmen in unserer Studie solche Merkmale ein, die sich auf die *Necessary Information (NI)* beziehen. Hierbei handelt es sich immer um jene Information, die benötigt wird, um ein Item korrekt zu bearbeiten (vgl. Buck & Tatsuoka, 1998). Zusätzlich zu den aus der Literatur gewonnenen Merkmalen ergänzten wir nach Gruppendiskussionen weitere schwierigkeitsbestimmende Aufgabenmerkmale, wie beispielsweise die automatisiert ermittelte

Überlappung der Stimulustexte mit den 500 häufigsten Wörter des Deutschen. Eine vollständige Auflistung sämtlicher von uns als potenziell relevant eingestufter schwierigkeitsbestimmender Merkmale ist an dieser Stelle aus Platzgründen leider nicht möglich. In Tabelle I.1 ist aber eine Auswahl schwierigkeitsbestimmender Merkmale für beide rezeptiven Sprachkompetenzen zusammengestellt, die einen Eindruck der berücksichtigten Bandbreite an Merkmalen geben kann.

Nachdem eine Definition und Beschreibung möglicher relevanter Merkmale vorgenommen wurde, ist es erforderlich, alle im jeweiligen Test eingesetzten Aufgaben und Items im Hinblick auf diese Merkmale einzustufen. Derartige Einschätzungen sollten von Fachexperten oder geschulten Bewertern vorgenommen werden. Hierbei werden die theoretisch angenommenen Ausprägungen oder Abstufungen der Merkmale in eine numerische Einschätzung überführt. Die Beurteilungen können quantitativ abgestuft oder kategorial gewählt sein. Wird für ein Merkmal lediglich sein Vorliegen beziehungsweise sein Fehlen eingeschätzt, so wählt man eine dichotome Kodierung mit den Ausprägungen „0 = Merkmal liegt nicht vor“ und „1 = Merkmal liegt vor“. Wird jedoch für ein Merkmal seine quantitative Ausprägung eingeschätzt, so werden hierfür Einschätzungen auf einer mehrstufigen Ratingskala abgegeben. Hierbei wird die Zuweisung von numerischen Einschätzungen zumeist so gewählt, dass sich ein positiver Zusammenhang zur empirischen Itemschwierigkeit ergibt.

Sind die schwierigkeitsbestimmenden Aufgaben- und Itemmerkmale definiert und für die vorliegenden Aufgaben kodiert worden, so können diese mit den empirisch ermittelten Itemschwierigkeiten in Zusammenhang gebracht werden.

Tabelle I.1: Schwierigkeitsbestimmende Merkmale für das Lese- und Hörverstehen

Auswahl schwierigkeitsbestimmender Aufgabenmerkmale, die sowohl für das Lese- als auch für das Hörverstehen als relevant identifiziert wurden:	Mittlere Satzlänge
	Durchschnittliche Silbenanzahl pro Wort
	Überlappung mit Grundwortschatz
	Verhältnis von Inhalts- und Funktionswörtern
	Redundante Repräsentation der Necessary Information (NI)
	Notwendigkeit von globaler vs. lokaler Kohärenzbildung
	Notwendigkeit von Vorwissen
	Itemtyp
	Komplexität des Items
Ergänzende schwierigkeitsbestimmende Aufgabenmerkmale, die für das Hörverstehen als relevant identifiziert wurden:	Aufnahmequalität
	Deutlichkeit der Aussprache
	Pausen zwischen Sätzen und Satzteilen
	Dialekt-, Regiolekt-, Soziolektfärbung
	Zahl der Sprecher/-innen
	Unterscheidbarkeit der Sprecherstimmen
	Überlappende Sprecherwechsel
	Sprechgeschwindigkeit der Necessary Information (NI)
	Sprechgeschwindigkeit für Umgebungsinformation der NI
	Hinweis darauf, dass NI folgen wird
	Deutliche Betonung der NI

I.3 Methode

I.3.1 Beschreibung der Personen- und Itemstichprobe

Die Pilotierung des Aufgabenpools zur Evaluierung der Bildungsstandards im Primarbereich (vgl. KMK, 2005) erfolgte im Rahmen der IGLU-Studie 2006. Im Frühjahr des Jahres 2007 fand die Normierung dieses Aufgabenpools im Zusammenhang mit der TIMSS-Studie statt. Die hier vorgestellten Daten entstammen beiden Studien. Die Schülerinnen und Schüler bearbeiteten zwei Mal 40 Minuten lang Aufgaben aus den verschiedenen Kompetenzbereichen der Fächer Deutsch und Mathematik. Hierbei nahmen aus jeder Schule je eine dritte und eine vierte Klasse teil. Neben vielen anderen Aufgaben wurden zehn literarische Lesetexte, acht Sachtexte sowie sechs Hörtexte eingesetzt. Die sechs Höraufgaben umfassen insgesamt 62 Testitems, davon sind 35 als geschlossen (*multiple choice*), 18 als halboffene und neun als offene Items einzustufen. Etwas mehr als die Hälfte dieser Items erfragen konkret gegebene Textinformationen, die restlichen Items erfordern eine Weiterverarbeitung von Textinformationen in Form von Inferenzen oder Interpretationen.

Zu den verwendeten Hörtexten zählen vertonte Märchen, ein Auszug aus einem vertonten Kinderroman, (mehrteilige) Radiosendungen für Kinder sowie eine Aufgabe zur auditiven Aufmerksamkeitsprüfung. Der Gehalt an authentischer Mündlichkeit variiert somit deutlich zwischen den verschiedenen Stimuli. Die sechs Höraufgaben wurden von insgesamt N = 3.757 Schülerinnen und Schülern der dritten und vierten Jahrgangsstufe bearbeitet. Die Stichprobe verteilt sich weitgehend balanciert auf die Pilotierungs- und Normierungsstudie sowie auf die dritte und vierte Jahrgangsstufe.

Da die Itementwicklung in unserer Studie nicht unter expliziter Berücksichtigung der schwierigkeitsbestimmenden Merkmale erfolgte, wurden alle eingesetzten Aufgaben und Items im Nachhinein im Hinblick auf die von uns als potenziell relevant identifizierten Merkmale von geschulten Ratern eingestuft. Die Einschätzungen erfolgten je nach Merkmal entweder dichotom (trifft zu/trifft nicht zu) oder auf einer mehrstufigen Ratingskala.

I.3.2 Grundlagen der Skalierung

Um die eingeschätzten Aufgabenmerkmale mit den empirisch ermittelten Itemschwierigkeiten in Verbindung setzen zu können, wählten wir ein zweischrittiges Verfahren. Zuerst führten wir eine

Skalierung der Daten zur Bestimmung der Itemschwierigkeiten durch, um anschließend in einer Regressionsanalyse die Beziehung zu den Aufgabenmerkmalen herzustellen.

Wie im Kontext großer Schulleistungsstudien üblich, wählten wir für die Skalierung der oben vorgestellten Daten ein Modell der probabilistischen Testtheorie (auch *Item Response Theory,* IRT, vgl. z. B. Embretson & Reise, 2000). Der entscheidende Vorteil von IRT-basierten Modellen ist die Möglichkeit, Schülerfähigkeiten und Itemschwierigkeiten auch im Rahmen eines Multi-Matrix-Designs auf einer gemeinsamen Skala abbilden zu können (vgl. De Boeck & Wilson, 2004). Da im Rahmen von Large-Scale-Assessments zumeist eine große Anzahl an Aufgaben und Items empirisch evaluiert werden soll, ist es nicht möglich, jedem Schüler und jeder Schülerin alle existierenden Aufgaben zur Bearbeitung vorzulegen. Vielmehr beschäftigte sich jedes Kind nur mit einer kleinen Auswahl von Aufgaben eines Tests. Diese Zuordnung bestimmter Aufgaben und Items zu Personen wird als (komplexes) Multi-Matrix-Design bezeichnet. Konkret wählten wir pro Kompetenzbereich, also für Lesen und Hören getrennt, ein eindimensionales Rasch-Modell, welches in der von uns eingesetzten Software *ConQuest 2.0* (Wu, Adams, Wilson & Haldane, 2007) implementiert ist.

I.4 Regressions- und Kommunalitätenanalysen

Um den Zusammenhang zwischen Aufgabenmerkmalen und Itemschwierigkeiten zu modellieren, wählten wir ein additives lineares Modell. Die Umsetzung dieses Modells erfolgte in einer multiplen linearer Regressionsanalyse. Diese Analysemethode findet in den Sozialwissenschaften verbreitet Anwendung und eröffnet die Möglichkeit, die Menge an Varianz zu bestimmen, die zwei oder mehr Prädiktorvariablen für eine einzelne Kriteriumsvariable aufklären können. Im hier vorliegenden Fall sind diese Kriteriumsvariablen die empirisch ermittelten Itemschwierigkeiten der Hör- sowie Leseverstehensaufgaben. Die Prädiktorvariablen sind die schwierigkeitsbestimmenden Merkmale auf Text- und Itemebene. Die empirischen Itemschwierigkeiten werden somit als gewichtete Summe der relevanten schwierigkeitsbestimmenden Aufgabenmerkmale modelliert. Entscheidend ist hierbei, inwieweit die Unterschiede in den Itemschwierigkeiten durch die Aufgabenmerkmale erklärt werden können. Der Determinationskoeffizient R^2 gibt den durch die Aufgabenmerkmale erklärten Varianzanteil in den Itemschwierigkeiten an. Die praktische Relevanz eines einzelnen Prädiktors wird üblicherweise über den standardisierten Regressionskoeffizienten β quantifiziert, auch wenn im Fall dichotomer Prädiktoren im Allgemeinen der unstandardisierte Koeffizient eine leichtere

Interpretierbarkeit besitzt. Wir verwenden ferner das LMG-Maß[2] der relativen Prädiktorrelevanz (vgl. Grömping, 2006, 2007).

In den von uns durchgeführten Regressionsanalysen haben wir zum einen Merkmale berücksichtigt, die nach den in der Literatur berichteten Befunden sowohl für das Hör- als auch für das Leseverstehen relevant sein sollten. Hierbei handelt es sich beispielsweise um die Explizitheit der *Necessary Information* oder auch um den Itemtyp. Zusätzlich haben wir sowohl für das Hör- als auch für das Leseverstehen solche Merkmale in die Regressionsanalysen einbezogen, die nur spezifisch relevant sind. Ein Beispiel wäre die Sprechgeschwindigkeit in der Umgebung der *Necessary Information*. Dieses Merkmal kann nur für Hörverstehenstexte sinnvoll eingesetzt werden.

Zusätzlich haben wir in unseren Analysen unterschieden, ob die schwierigkeitsbestimmenden Merkmale auf Item- oder auf Textebene wirken. Die Frage, wie oft die Information, die für die Beantwortung eines Items benötigt wird, im Text vorkommt, ist ein Merkmal, welches die Itemebene betrifft. Die mittlere Satzlänge eines Textes lässt sich hingegen nur für den gesamten Text bestimmen und stellt somit ein Merkmal auf Textebene dar.

Da wir für den Bereich des Hörverstehens nur auf die Daten zu sechs Aufgaben zurückgreifen konnten, haben wir, um eine Selektion in den betrachteten Items zu vermeiden, vor Berechnung der Regressionsanalysen fehlende Werte imputiert (Lüdtke, Robitzsch, Trautwein & Köller, 2007). Hierbei ist zu betonen, dass dieser Ersetzungsprozess fehlender Merkmale nicht der „Datenerfindung" (Rost, 2007) dient, sondern das Analyseproblem unter Nutzung aller verfügbaren Informationen behandelt. Eine ausschließliche Betrachtung von Items mit vollständigen Daten wird im Allgemeinen zu einer nichtrepräsentativen Itemstichprobe führen und es besteht die Gefahr, verzerrte Schätzungen zu erhalten.

Prädiktorvariablen in multiplen Regressionen sind oftmals miteinander korreliert, was auch unter dem Begriff der Kollinearität diskutiert wird und die Interpretation der Befunde deutlich erschweren kann (vgl. Pedhazur, 1997; Zientek & Thompson, 2006). Liegen korrelierte Prädiktoren vor, so genügt es nicht, ausschließlich die β-Gewichte der Prädiktoren zu interpretieren, um ihren Stellenwert bei der Vorhersage des Kriteriums zu ermitteln. Zusätzlich sollte im Fall korrelierter Prädiktoren zusätzlich die Korrelation des jeweiligen Prädiktors mit dem Kriterium ermittelt werden. Auch das von uns verwendete LMG-Maß der relativen Prädiktorrelevanz dient dazu, dem Problem korrelierter Prädiktoren Rechnung zu tragen. Bildet man die Summe aller LMG-Koeffizienten, erhält man den Anteil der durch alle Prädiktoren

[2] Die Abkürzung „LMG" steht für die Namen der Autoren Lindeman, Merenda und Gold, die dieses Maß vorgeschlagen haben.

aufgeklärten Varianz, der dem Determinationskoeffizienten R^2 entspricht. Im Gegensatz zu den standardisierten Regressionskoeffizienten wird also das gesamte R^2 in auf die Prädiktoren zurückführbare Einzelanteile zerlegt. Falls alle Prädiktoren unkorreliert sind, entspricht allerdings die Summe der quadrierten standardisierten Regressionskoeffizienten ebenfalls R^2.

Ein weiterer viel versprechender Zugang, um den Beitrag jedes Prädiktors zur Aufklärung des Kriteriums einzuschätzen, stellt die eher selten eingesetzte Kommunalitätenanalyse (*Commonality Analysis*) dar (vgl. z. B. Beaton, 1973). In der Kommunalitätenanalyse wird der Anteil aufgeklärter Varianz R^2 so zerlegt, dass für einzelne Prädiktoren spezifische Anteile (*Unique*) von den durch alle möglichen Kombinationen an Prädiktoren gemeinsam aufgeklärten Anteilen (*Common*) unterschieden werden können.[3] Veranschaulicht wird die Idee der Kommunalitätenanalyse in Abbildung I.1, in der zwei korrelierte Prädiktorvariablen X_1 und X_2 zur Erklärung der Kriteriumsvariable Y eingesetzt werden. Es wird deutlich, dass beide Prädiktorvariablen einen jeweils spezifischen Anteil an Varianz aufklären können, dieser ist in der Grafik mit U für *Unique* gekennzeichnet (U_{X1}, U_{X2}). Darüber hinaus gibt es aber auch einen Anteil gemeinsamer Varianz, der überlappend durch beide Prädiktoren erfasst wird. Dieser Anteil ist mit C für *Common* bezeichnet (C_{X12}). In der unten stehenden Abbildung I.1 entspricht dabei $U_{X1}+C_{X12}$ dem Quadrat der Korrelation von Y mit X_1, so dass der Anteil von U_{X1} als für die Variable X_1 spezifischer Varianzanteil am Kriterium Y interpretiert werden kann.

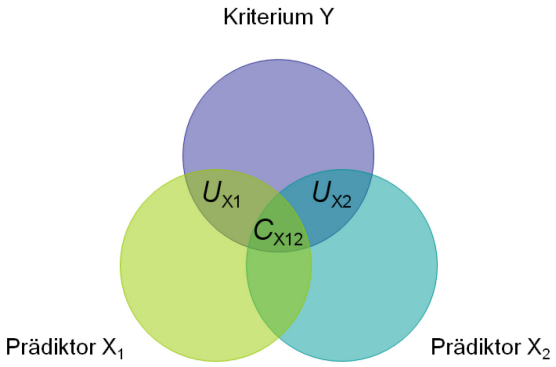

Abbildung I.1: Schematisches Venn-Diagramm zur Veranschaulichung von spezifischer (Uniquer) Varianz der Prädiktoren X_1 und X_2 (U_{X1}, U_{X2}) sowie der gemeinsam erklärten Varianz (Common, $C_{X1, X2}$)

Alle Analysen wurden in der Software R (R Development Core Team, 2007) unter Zuhilfenahme der Pakete „mice“ für die Imputation fehlender Daten (Van Buuren & Oudshoorn, 2007), „relaimpo“ für die Berechnung der relativen Wichtigkeit der Prädiktorvariablen (Grömping,

[3] Dieses Vorgehen entspricht nicht der Analyse von Interaktionseffekten in Varianzanalysen

2006, 2007) sowie einer R-Routine für die Kommunalitätenanalyse (Nimon, Lewis, Kane & Haynes, 2008) durchgeführt.

I.5 Empirische Befunde

I.5.1 Ergebnisse zum Hörverstehen

In Tabelle I.2 sind die Befunde einer multiplen Regressionsanalyse zur Vorhersage der Itemschwierigkeit durch Merkmale auf Item- und Textebene dargestellt. Um den Stellenwert eines bestimmten Prädiktors einzuschätzen, können nun der standardisierte Regressionskoeffizient β, das LMG-Maß der relativen Prädiktorrelevanz, welches auch Kollinearitäten berücksichtigt, und die Korrelation r des Prädiktors mit dem Kriterium betrachtet werden. Greift man beispielhaft das Merkmal „Redundante Präsentation der NI" heraus, so wird deutlich, dass es sich hierbei um einen Prädiktor mit signifikantem Regressionsgewicht handelt (β = .37, p = .01), der auch deutlich mit dem Kriterium der Itemschwierigkeit korreliert ist (r = .37). Auch der hohe LMG-Wert von .08 weist auf die hohe Relevanz dieses Prädiktors hin, die auch aus theoretischer Sicht unmittelbar einsichtig ist. Nicht unerwähnt bleiben soll, dass die Mehrzahl der Variablen nicht als signifikante Prädiktoren identifiziert werden kann. Dies soll hier jedoch nicht im Vordergrund stehen, da wir vielmehr an der relativen Bedeutung der Prädiktoren interessiert sind.

Betrachtet man ausschließlich diejenigen in Tabelle I.2 aufgeführten Merkmale, die auf Itemebene ansetzen, so erreicht man für die Itemschwierigkeit der Hörverstehensaufgaben eine Varianzaufklärung von 24 %. Nimmt man jene Merkmale hinzu, die auf Textebene ansetzen, so steigt die Varianzaufklärung auf R^2 = .49.

Unterscheidet man nicht zwischen Merkmalen von Item- und Textebene, sondern zwischen allgemeinen Merkmalen, die sowohl für das Hör- als auch für das Leseverstehen relevant sind, so erhält man eine Varianzaufklärung von 31 %. Die Hinzunahme der hörspezifischen Merkmale, die in Tabelle I.2 grau unterlegt sind, erhöht die Varianzaufklärung also um weitere 18 %.

Tabelle I.2: Ergebnisse einer multiplen Regression zur Vorhersage der Itemschwierigkeit durch Merkmale auf Item- und Textebene sowie Korrelation mit dem Kriterium

Merkmalsebene	Merkmal	β	p	LMG	r
Itemebene	Explizitheit der *NI*	.13	.51	.00	.03
	Redundante Präsentation der *NI*	.37	.01	.08	.37
	Hintergrundwissen für Itembeantwortung notwendig	.01	.96	.01	.05
	Inferenzbildung für Itembeantwortung notwendig	.13	.52	.00	.06
	Itemtyp: halboffen	.14	.32	.01	.01
	Itemtyp: offen	.06	.70	.00	.08
	Sprechgeschwindigkeit in der Umgebung der *NI*	.32	.30	.02	.27
	Anzahl notwendiger Sprechakte für *NI*	.13	.44	.05	.37
	Dialekt-/Regiolekt- und Soziolektfärbung	.28	.13	.02	.19
Textebene	Mittlere Satzlänge	.00	.99	.01	.03
	Mittlere Wortlänge	.41	.15	.02	.19
	Prozentuale Überlappung mit häufigen Wörtern	.99	.16	.04	.25
	Lebensweltbezug	.47	.05	.14	.49
	Zahl der Sprecher/innen im Text	.90	.12	.09	.40

Anmerkung: β = standardisierter Regressionskoeffizient; p = p-Wert zur Kennzeichnung der Signifikanz; LMG = Maß der Relevanz des Prädiktors, die Summe der LMG-Koeffizienten ergibt die aufgeklärte Varianz R^2; r = Korrelation des Prädiktors mit dem Kriterium; grau hinterlegte Merkmale gelten als spezifisch für das Hörverstehen

Zusätzliche Kommunalitätenanalysen für die Merkmale auf Itemebene ergaben, dass die Merkmale „Redundante Präsentation der NI“ und „Anzahl notwendiger Sprechakte für NI“ 32 % beziehungsweise 48 % der mit der Kriteriumsvariablen „Itemschwierigkeit“ geteilten Varianz spezifisch aufklären, während die restlichen Varianzanteile mit anderen Prädiktoren konfundiert sind. Lediglich für das Merkmal „Sprechgeschwindigkeit in der Umgebung der NI“ zeigt sich aufgrund eines spezifischen Varianzanteils von 4 %, dass der überwiegende Teil der Varianz nicht spezifisch, sondern überlappend mit anderen Merkmalen aufgeklärt wird.

I.5.2 Vergleichende Ergebnisse zum Leseverstehen

Verwendet man in einer Regression für das Leseverstehen die in Tabelle I.3 dargestellten Merkmale, die sowohl für die Vorhersage der Itemschwierigkeit in Hör- als auch in Leseverstehensaufgaben eingesetzt werden können, erreicht man eine Varianzaufklärung von R^2 = .29, die also in absolut vergleichbarer Höhe wie für das Hörverstehen liegt.

Auch für das Leseverstehen können jedoch spezifische Merkmale identifiziert werden. Für einzelne Aufgaben, die als Stimuli literarische Texte verwenden, kann unter Berücksichtigung zusätzlicher Merkmale, wie der Prozentanteil wörtlicher Rede oder die Anzahl von Vergleichen und Metaphern, eine ähnlich hohe Varianzaufklärungen wie für das Hörverstehen erreicht werden.

Tabelle I.3: Ergebnisse einer multiplen Regression zur Vorhersage der Itemschwierigkeit durch Merkmale auf Item- und Textebene sowie Korrelation mit dem Kriterium

Merkmalsebene	Merkmal	β	p	LMG	r
Itemebene	Explizitheit der *NI*	.24	.06	.06	.31
	Redundante Präsentation der *NI*	.20	.04	.03	.16
	Hintergrundwissen für Itembeantwortung notwendig	.16	.19	.05	.33
	Inferenzbildung für Itembeantwortung notwendig	.05	.60	.02	.24
	Itemtyp: halboffen	.24	.01	.03	.06
	Itemtyp: offen	.29	.00	.06	.27
Textebene	Mittlere Satzlänge	.04	.71	.01	.13
	Mittlere Wortlänge	.38	.32	.01	.05
	Prozentuale Überlappung mit häufigen Wörtern	.24	.22	.01	.11
	Lebensweltbezug	.10	.40	.00	.05

Anmerkung: β = standardisierter Regressionskoeffizient; p = p-Wert zur Kennzeichnung der Signifikanz; LMG = Maß der Relevanz des Prädiktors, die Summe der LMG-Koeffizienten ergibt die aufgeklärte Varianz R^2;
r = Korrelation des Prädiktors mit dem Kriterium

I.6 Diskussion und Ausblick

Die vorgestellten Befunde zeigen deutlich, dass sowohl für das Hör- als auch für das Leseverstehen allgemeine Aufgabenmerkmale, die sich auf das Textverstehen beziehen, bei der Erklärung der Itemschwierigkeit einen wichtigen Stellenwert einnehmen. Um größere Teile der Varianz in der Itemschwierigkeit erklären zu können, sind jedoch Merkmale nötig, die spezifische Eigenschaften des Hör- beziehungsweise Leseverstehens beschreiben. Diese Befunde sind erwartungskonform, da in der theoretischen Betrachtung der beiden rezeptiven Sprachkompetenzen betont wird, dass die kognitiven Informationsverarbeitungsprozesse weitreichende Ähnlichkeiten aufweisen, auf der Ebene der Rezeption jedoch die verschiedenen Inputstimuli (visuell vs. auditiv) berücksichtigt werden müssen. Die Betrachtung schwierigkeitsbestimmender Aufgabenmerkmale kann daher als ein Erfolg versprechender empirischer Zugang verstanden werden, um theoretische Überlegungen zur Unterscheidung des Hör- und Leseverstehens zu bestätigen.

Problematisch bei der Arbeit mit schwierigkeitsbestimmenden Aufgabenmerkmalen ist jedoch der generelle Umstand, dass nicht alle Merkmale in gleicher Weise für die untersuchten Items und Texte sinnvoll und zutreffend sind. Die Ergebnisse zum Leseverstehen verdeutlichen beispielsweise, dass einige Merkmale nur textsortenspezifisch modelliert werden können, da sie nur innerhalb einer bestimmten Textsorte bedeutsam sind. Dies betrifft etwa das Merkmal „Anzahl der Vergleiche und Metaphern“, welches in der vorliegenden Studie für literarische Texte eine deutliche Aussagekraft besitzt, für Sachtexte jedoch einen viel geringeren Stellenwert einnimmt. Für die überwiegende Anzahl der von uns eingesetzten Sachtexte können kaum Vergleiche oder Metaphern identifiziert werden, weshalb für Sachtexte im Hinblick auf dieses Merkmal eine stark eingeschränkte Varianz vorliegt. Es scheint daher angezeigt, in weiterführenden Studien auch Interaktionseffekte zwischen Merkmalen und der Textsorte zu berücksichtigen.

Möglicherweise entsteht bei der Betrachtung der Befunde der Eindruck, dass die berichteten Varianzaufklärungen nicht zufriedenstellen können. Arbeitet man mit einer eher geringen Itemzahl und einer großen Zahl an theoretisch relevanten Merkmalen, besteht immer die Gefahr, nicht alle praktisch relevanten Merkmale identifizieren zu können, so dass eine Varianzaufklärung von mehr als 50 % unrealistisch erscheint. Nur bei einer experimentellen Itemgenerierung, die ganz gezielt bestimmte schwierigkeitsbestimmende Merkmale manipuliert und die Ausprägung aller übrigen Merkmale unverändert beibehält, könnte eine höhere Varianzaufklärung erwartet werden. Werden die Items, wie im hier vorliegenden Fall, nicht experimentell entwickelt, bestehen immer Konfundierungen und eine vollständige Aufklärung der Itemschwierigkeit wird nicht zu erreichen sein. Daraus folgt allerdings, dass eine kausale

Interpretation von gleichzeitiger Betrachtung von Regressionskoeffizienten für Prädiktoren im Regressionsmodell (auch bezüglich der Relevanz dieser) nicht möglich ist und die Ergebnisse nur deskriptiv zu bewerten sind (Morgan & Winship, 2007). Vor diesem Hintergrund ist ein auf nichtexperimenteller Itemgenerierung basierendes Regressionsmodell mit schwierigkeitsbestimmenden Itemmerkmalen, das zur Definition und/oder Beschreibung von Kompetenzstufenmodellen dient, kritisch zu bewerten (vgl. Hartig, 2007, für das Vorgehen in DESI). Zentraler Hintergrund dieses Beitrags ist jedoch auch, für den zukünftigen Prozess der Aufgabenentwicklung bestimmte Merkmale, die sich als relevant erwiesen haben, systematisch so zu manipulieren, dass zielgerichtet Items einer intendierten Schwierigkeit entstehen. Hierbei würde es sich dann um eine (partiell) experimentelle Itemgenerierung handeln.

Die größte Einschränkung der hier berichteten Befunde ergibt sich aus der Beurteilung der schwierigkeitsbestimmenden Merkmale in den Aufgaben. Diese Beurteilungen wurden von Studierenden im Rahmen der Anfertigung ihrer Masterarbeiten vorgenommen. Obwohl alle Studierenden intensiv geschult und im Verlauf des Ratingprozesses von Experten beraten wurden, ist es dennoch erforderlich, in weiterführenden Studien zusätzliche Einschätzungen zu generieren, um die Übereinstimmungen der Rater bei der Beurteilung der Merkmale ermitteln und die hier vorgestellten Befunde auf diese Weise absichern und verallgemeinern zu können.

Einen wichtigen Ansatzpunkt zur Fortführung der Forschungsarbeiten zu schwierigkeitsbestimmenden Merkmalen stellt der Einsatz elaborierterer Analysemethoden dar. Zunächst wäre eine Übertragung des regressionsanalytischen Zugangs auf den Bereich der probabilistischen Testtheorie in Form des *Linearen Logistischen Testmodells* (LLTM) von Fischer (1973) denkbar. Bei Anwendung des LLTM wird das hier eingesetzte zweischrittige Vorgehen, bei dem zunächst mittels einer Rasch-Skalierung die empirischen Itemschwierigkeiten ermittelt werden müssen und diese erst anschließend in einer Regression mit den eingeschätzten Aufgabenmerkmalen in Verbindung gebracht werden können, überflüssig. Wesentlicher Unterschied zwischen dem zweischrittigen Vorgehen und dem LLTM-Modell ist der, dass in letzterem Fall die Itemschwierigkeiten ausschließlich durch die Aufgabenmerkmale erklärt werden und kein Residuum angenommen wird. Bei der Bestimmung standardisierter Regressionskoeffizienten ist zu beachten, dass bei einem zweistufigen Verfahren infolge der Messfehlerbehaftetheit der Itemschwierigkeiten im Allgemeinen eine Unterschätzung der Koeffizienten erfolgt.

In folgenden Analysen ist bei hinreichender Anzahl von Hörstimuli ferner die Item- von der Textebene zu separieren (vgl. Ozuru, Rowe, O'Reilly & McNamara, 2008, für ein Beispiel für das Leseverstehen). Durch diesen Ansatz lassen sich auf Text- und Itemebene zurückführbare

Varianzanteile hinsichtlich der Relevanz auf den einzelnen Ebenen interpretieren, während dies im linearen Regressionsmodell konfundiert bleibt.

Einen weiteren Ansatzpunkt stellen mehrdimensionale IRT-Modelle dar. Die Frage, ob Items als Indikatoren einer einzigen, klar separierbaren kognitiven Fähigkeit verstanden werden können oder ob die Lösung eines Items nicht immer ein komplexes Zusammenspiel einer Vielzahl kognitiver Fähigkeiten erfordert, ist zentral, um eine zutreffende Modellierung der Daten und somit eine adäquate Abbildung der dimensionalen Kompetenzstruktur vornehmen zu können. In der einschlägigen Literatur wird davon ausgegangen, dass zum Lösen eines Items in aller Regel ein Zusammenwirken mehrerer Kompetenzen benötigt wird und Items daher zumeist mehr als eine einzige, singuläre Fähigkeit überprüfen (vgl. Reckase, Ackerman & Carlson, 1988). Eine Möglichkeit, diese Überlegung methodisch umzusetzen, sind mehrdimensionale IRT-Modelle (MIRT-Modelle). In diesen Modellen wird die Struktur kategorialer, beobachtbarer Variablen in multiplen, kontinuierlichen, latenten Dimensionen repräsentiert. Die Anwendung von MIRT-Modellen für den hier relevanten Kontext erfolgte in jüngster Zeit zum Beispiel durch Hartig und Höhler (2008). Die Autoren konnten zeigen, dass die Antworten zu Hör- und Leseverstehensitems der DESI-Studie für Englisch als Fremdsprache in zweidimensionalen IRT-Modellen repräsentiert werden können. Vergleichbare Modelle prüfen wir für das Lese- und Hörverstehen im Grundschulkontext.

I.7 Literatur

Beaton, A. E. (1973). Commonality. (ERIC Document Reproduction Service *No. ED 111 829*)

Buck, G. (2001). *Assessing listening*. New York: Cambridge University Press.

Buck, G. & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing*, *15*, 119-157.

De Boeck, P. & Wilson, M. (2004). *Explanatory item response models*. New York: Springer.

Embretson, S. E. & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum Publishers.

Faerch, C. & Kasper, G. (1986). The role of comprehension in second language learning. *Applied Linguistics*, *7,* 257-274.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37, 359-374.

Freedle, R. & Kostin, I. (1993). The prediction of TOEFL reading item difficulty: implications for construct validity. *Language Testing*, *10*, 133-170.

Grömping, U. (2006). Relative importance for linear regression in R: The package relaimpo. *Journal of Statistical Software*, *17(1)*, 1-27.

Grömping, U. (2007). Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician*, *61*, 139-147.

Grotjahn, R. (2000). *Determinanten der Schwierigkeit von Leseverstehensaufgaben. Theoretische Grundlagen und Konsequenzen für die Entwicklung des TESTDAF*. München: Gilde-Verlag.

Grotjahn, R. (2005). Testen und Bewerten des Hörverstehens. In M. Ó. Dúill, R. Zahn & K. D. C. Höppner (Hrsg.), *Zusammenarbeiten: Eine Festschrift für Bernd Voss* (S. 115-144). Bochum: AKS-Verlag.

Hartig, J. (2007). Skalierung und Definition von Kompetenzniveaus. In B. Beck & E. Klieme (Hrsg.), *Sprachliche Kompetenzen – Konzepte und Messung* (S. 83-99). Weinheim: Beltz.

Hartig, J. & Höhler, J. (2008). Representation of competencies in multidimensional IRT models with within-item and between-item multidimensionality. *Zeitschrift für Psychologie*, *216,* 89-101.

KMK (2005). siehe Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2005).

Köster, J. (2005). Wodurch wird ein Test schwierig? Ein Text für die Fachkonferenz. *Deutschunterricht, 58(5),* 34-39.

Kürschner, C. & Schnotz, W. (2008). Das Verhältnis gesprochener und geschriebener Sprache bei der Konstruktion mentaler Repräsentationen. *Psychologische Rundschau*, *59*, 139-149.

Leucht, M., Retelsdorf, J., Möller, J. & Köller, O. (2010). Zur Dimensionalität rezeptiver englischsprachiger Kompetenzen. *Zeitschrift für Pädagogische Psychologie*, *24,* 123-138.

Lüdtke, O., Robitzsch, A., Trautwein, U. & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung: Probleme und Lösungen. *Psychologische Rundschau*, *58,* 103-117.

Lynch, T. (1998). Theoretical perspective on listening. *Annual Review of Applied Linguistics*, *18,* 3-19.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*, 5-11.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23,* 13-23.

Morgan, S. L. & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research.* Cambridge: Cambridge University Press.

Nimon, K., Lewis, M., Kane, R. & Haynes, R. M. (2008). An R package to compute commonality coefficients in the multiple regression case: An introduction to the package and a practical example. *Behavior Research Methods*, *40,* 457-466.

Nold, G. & Rossa, H. (2007a). Hörverstehen. In E. Klieme & B. Beck (Hrsg.), *Sprachliche Kompetenzen – Konzepte und Messung* (S. 174-192). Weinheim: Beltz.

Nold, G. & Rossa, H. (2007b). Leseverstehen. In E. Klieme & B. Beck (Hrsg.), *Sprachliche Kompetenzen – Konzepte und Messung* (S. 193-207). Weinheim: Beltz.

Ozuru, Y., Rowe, M., O'Reilly, T. & McNamara, D. S. (2008). Where's the difficulty in standardized reading tests: the passage or the question? *Behavior Research Methods*, *40,* 1001-1015.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research.* Orlando, FL: Harcourt Brace.

R Development Core Team (2007). *R: A language and environment for statistical computing*. Wien: R Foundation for Statistical Computing.

Reckase, M. D., Ackerman, T. A. & Carlson, J. E. (1988). Building an unidimensional test using multidimensional items. *Journal of Educational Measurement*, *25*, 193-203.

Rost, D. H. (2007). *Interpretation und Bewertung pädagogisch-psychologischer Studien*. Weinheim: Beltz.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2005). *Bildungsstandards im Fach Deutsch für den Primarbereich (Jahrgangsstufe 4) – Beschluss vom 15.10.2004*. München: Wolters Kluwer.

Van Buuren, S. & Oudshoorn, C. G. M. (2007). mice: multivariate imputation by chained equations. R package version 1.16. Zugriff am 10.10.2011 unter http://web.inter.nl.net/users/S.van.Buuren/mi/hmtl/mice.htm

Willenberg, H. (2007). Lesen. In E. Klieme & B. Beck (Hrsg.), *Sprachliche Kompetenzen – Konzepte und Messung* (S. 107-117). Weinheim: Beltz.

Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. A. (2007). *ACER Conquest 2.0*. Melbourne: The Australian Council for Educational Research.

Zientek, L. R. & Thompson, B. (2006). Commonality analysis: Partitioning variance to facilitate better understanding of data. *Journal of Early Intervention*, *28*, 299-307.

3.3 Kompetenzbereich II – Schreiben

Der Kompetenzbereich Schreiben wird in der vorliegenden Arbeit in zwei empirischen Beiträgen behandelt. Hiervon widmet sich einer dem Bereich des freien Schreibens (Abschnitt II), der andere dem Bereich der orthografischen Kompetenz (Abschnitt III).

Da in den Beiträgen jeweils eine ausführliche theoretische Einbettung der betrachteten Kompetenzen enthalten ist, soll eine solche an dieser Stelle nicht zusätzlich thematisiert werden. Vielmehr sollen einige allgemeine Überlegungen zu Problemen und Herausforderungen der Testdiagnostik im Kompetenzbereich Schreiben zu den empirischen Beiträgen überleiten.

Der Kompetenzbereich Schreiben umfasst in den Bildungsstandards für das Fach Deutsch im Primarbereich (KMK, 2005a) ebenso wie für die Sekundarstufe I (KMK, 2004, 2005b) sowohl Aspekte des freien Schreibens beziehungsweise der Textproduktion als auch der orthografischen Kompetenz (vgl. Abschnitt 3.3.2).

3.3.1 Texte verfassen

Im Primarbereich lassen sich auf Seiten des freien Schreibens unter der Überschrift *Texte verfassen* die Komponenten *Texte planen*, *Texte schreiben* sowie *Texte überarbeiten* identifizieren (KMK, 2005a, S. 11). Den hierbei genannten Einzelstandards liegt die Überlegung zugrunde, dass Schreiben als Prozess zu verstehen ist, der sich aus den Phasen der Planung, Ausführung und Überwachung beziehungsweise Überarbeitung zusammensetzt. In diesem Sinne sind auch zahlreiche Kompetenzmodelle zum Schreiben freier Texte als Prozessmodelle angelegt (vgl. bspw. Bereiter & Scardamalia, 1987; Flower & Hayes, 1981; Hayes & Flower, 1980; Kellogg, 1996).

Bei der Entwicklung der Testinstrumente zur Evaluierung der Bildungsstandards für das Fach Deutsch im Primarbereich wurden intensive Bemühungen unternommen, Testaufgaben zu allen drei Phasen des Schreibprozesses zu konstruieren. Dies erwies sich durch die Limitationen der Testadministration als ausgesprochen schwierig.

Probeweise in den Test integrierte Aufgabenstellungen, die darauf abzielten, vor dem Verfassen des eigentlichen Textes von den Kindern stichpunktartig notieren zu lassen, wie diese das Schreiben des Textes planen, erwiesen sich als zeitaufwändig, inhaltlich kaum ergiebig und konnten den testmethodischen Gütekriterien nicht genügen. Der Prozessschritt der Textplanung kann bei Grundschulkindern allerdings auch nicht über ein vollständiges schriftliches Gedankenprotokoll erfasst werden, da dies für Kinder dieser Altersstufe bei weitem zu schreibintensiv wäre. Durch Gruppentestungen und die ausschließliche Verwendung von Paper-

Pencil-Tests war im Rahmen dieser Studie aber auch keine Aufzeichnung von lautem Denken möglich.

Aus den genannten Gründen beschränkten sich die in der Pilotierungsstudie eingesetzten Testaufgaben auf die Prozessschritte des Verfassens von Texten sowie der Überarbeitung. Bei letztgenanntem Prozessschritt wurden sowohl Aufgaben erprobt, bei denen die Schülerinnen und Schüler kurze eigene Texte überarbeiten sollten als auch Aufgaben, bei denen bereits vorliegende fremde Texte hinsichtlich bestimmter Kriterien überarbeitet werden sollten. Beide Formen der Überarbeitung werden durch die Bildungsstandards für das Fach Deutsch im Primarbereich abgedeckt. Leider erwiesen sich auch diese Aufgabentypen als für die Normierungsstudie ungeeignet. Neben der erforderlichen Testzeit für das Verfassen des Originaltextes und die sich anschließende Überarbeitung ergaben sich auch Probleme mit der testadministrativen Umsetzung, da die Kinder aus Gründen der Lesbarkeit und späteren Auswertbarkeit die Überarbeitungen nicht in ihrem Originaltext vornehmen konnten und das komplette Neuschreiben des Textes oftmals zu einem bloßen Abschreiben in schönerer Schreibschrift führte. Das eigentliche Hauptproblem besteht aber darin, dass Umfang und Qualität von Überarbeitungen wesentlich durch die Qualität des Ausgangstextes determiniert werden. Schülerinnen und Schüler schreiben sehr unterschiedlich lange und unterschiedlich komplexe Texte, dementsprechend gibt es mehr oder weniger Überarbeitungsspielraum und -bedarf. Diese Heterogenität in den Ausgangstexten und damit in den Überarbeitungen macht eine Beurteilung dieser Kompetenz ausgesprochen schwierig und führt dazu, dass man sich darauf beschränken müsste, festzuhalten, ob überhaupt in irgendeiner Form überarbeitet wurde, was sowohl aus fachdidaktischer wie auch aus psychometrischer Perspektive ausgesprochen unbefriedigend wäre. Ferner ist relevant, dass eine Überarbeitung von Texten in aller Regel durch Impulse von außen, beispielsweise durch die Lehrkraft oder Mitschüler, angestoßen wird. Diese Impulsgebung war während der Testung nicht möglich, was dazu führte, dass die Kinder Probleme hatten, Überarbeitungsbedarf in ihren Texten zu identifizieren und die Aufgabe somit nicht bearbeiten konnten. Bei der Überarbeitung fremder Texte wiederum führten die detaillierten Hinweise in der Aufgabenstellung, nach welchen Kriterien die vorgegebenen Texte überarbeitet werden sollten, dazu, dass sich sämtliche Schülerantworten sehr ähnlich waren und kaum variierten. Dieser Mangel an Varianz stellt testmethodisch jedoch ein großes Problem dar, da das Ziel der Testung ist, mit Hilfe der Aufgabenlösungen zwischen unterschiedlich kompetenten Schülerinnen und Schülern zu differenzieren. Wenn aber Kinder mit „objektiv“, also tatsächlich unterschiedlichen Kompetenzständen Aufgaben in sehr ähnlicher Weise (richtig oder falsch) bearbeiten, ist eine solche Differenzierung nicht möglich.

Aus den dargestellten Gründen beschränkte sich die Überprüfung der Schreibkompetenz im Rahmen der Evaluierung der Bildungsstandards somit auf das Verfassen von Texten. Trotz anfänglicher Bemühungen, der Prozessqualität des Schreibens auch in den Testinstrumenten Rechnung zu tragen, konnte also letztlich nur *eine* Phase des Schreibprozesses abgebildet werden. Dies führte dazu, dass bislang nicht der in den Bildungsstandards angelegte Schreibprozess sondern folglich nur das *Schreibprodukt* Gegenstand der Diagnostik ist.

Wie Schreibprodukte von Schülerinnen und Schülern bewertet werden können, ist zentraler Gegenstand des nachfolgenden empirischen Beitrags.

II. Beitrag 2: Aspekte der Kodierung von Schreibaufgaben

Autoren:
Katrin Böhme / Albert Bremerich-Vos / Alexander Robitzsch

Erschienen in:
D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.) (2009). *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 290-329). Weinheim: Beltz.

Dieser Beitrag verfolgt das Ziel, fachdidaktische Überlegungen zum Konstrukt der Schreibkompetenz mit diagnostischen Fragen der Operationalisierung und Bewertung von Schreibprodukten zu verknüpfen. Aus didaktischer Sicht steht hier die Frage nach den Komponenten dieser hochkomplexen Kompetenz im Mittelpunkt. Aus diagnostischer Perspektive handelt es sich hierbei um die Problematik der Dimensionalität des Konstrukts, die sich beispielsweise in verschiedenen Varianten der Kodierung der Schreibprodukte niederschlägt. Darüber hinaus sind Aspekte der Interraterreliabilität, also der Übereinstimmung zwischen verschiedenen Bewertern bei der Beurteilung von Schreibprodukten, zentral.

Um einen Vergleich holistischer und analytischer Kodierungen unter besonderer Berücksichtigung der Interraterreliabilität leisten zu können, werden Befunde für eine narrative Schreibaufgabe vorgestellt. Im Hinblick auf die erreichten Beurteilerübereinstimmungen zeigt sich für dichotome Variablen der analytischen Kodierstrategie ein befriedigendes Befundmuster. Vereinzelt ergeben sich Probleme aufgrund stark ungleicher Kategorienbesetzungen. Mehrstufige analytische Variablen, die zumeist dem allgemein sprachlichen Bereich entstammen, können allerdings mehrheitlich nicht mit ausreichender Reliabilität gemessen werden. Dieser Befund lässt sich in Multi-Trait-Multi-Method-Modellierungen unter Rückgriff auf Strukturgleichungsmodelle replizieren. Die für vier- und sechsstufige holistische Variablen ermittelten Intraklassenkorrelationen, die als mittlere Korrelation zweier Rater interpretiert werden können, betragen .67 bis .75, was auch im Vergleich mit der internationalen Literatur als Erfolg gedeutet werden kann.

Beim Vergleich der holistischen und analytischen Kodierstrategie bei der Beurteilung des Inhalts eines narrativen Textes zeigt sich ein hoher korrelativer Zusammenhang, der je nach gewähltem Raterdesign im Bereich von $r = .80$ bis .90 liegt. Eine analytische Summenvariable für die Bewertung des Inhalts misst im hier diskutierten Beispiel geringfügig reliabler als eine holistische Einschätzung des Inhalts.

Betrachtet man die Zusammenhänge der (hinreichend reliabel messbaren) analytischen Sprachvariablen untereinander, ergeben sich durchgängig moderate positive Korrelationen. Den höchsten Zusammenhang mit dem holistischen Gesamteindruck zeigt die Variable Wortschatz, sie ist auch in einer linearen Regression der beste Prädiktor. Alle Variablen des allgemeinen Sprachbereichs der analytischen Kodierung weisen deutlich geringere Halo-Effekte als holistische Einschätzungen der Schülertexte auf.

Befunde zur Dimensionalität des Konstrukts der Schreibkompetenz lassen einen engen Zusammenhang zwischen stilistischen und inhaltlichen Aspekten erkennen. Die Komponente der sprachlichen Richtigkeit (Rechtschreibung und Grammatik) nimmt eine Sonderstellung ein

und lässt sich nicht ohne Weiteres als Bestandteil eines eindimensionalen Konstrukts fassen. Dies steht im Einklang mit Befunden für den deutschsprachigen Raum (vgl. Neumann, 2007).

Inwieweit diese Ergebnisse Implikationen für die Verwendung analytischer und holistischer Kodierstrategien bergen, hängt davon ab, welche Ziele und Fragestellungen in den jeweiligen Studien verfolgt werden. Hierbei sollte zwischen *Large-Scale-Assessments* mit dem Ziel des Bildungsmonitorings und auf Individualdiagnostik zielenden, didaktisch motivierten Studien unterschieden werden.

II.1 Überblick

In diesem Beitrag geht es uns um die Klärung der Frage, wie Schreibkompetenz im Grundschulalter unter Berücksichtigung der Vorgaben der Bildungsstandards für den Primarbereich (vgl. KMK, 2005) im Kontext des Large-Scale-Assessments definiert und in eine aussagekräftige Leistungsmessung überführt werden kann. Hierfür werden in Abschnitt II.2 zunächst fachdidaktische Überlegungen zum Konstrukt der Schreibkompetenz und ihrer Entwicklung zusammengefasst. In Abschnitt II.3 setzen wir uns mit der Beschreibung der Schreibkompetenz in den Bildungsstandards und Möglichkeiten einer Operationalisierung des Konstrukts im Rahmen großer Schulleistungstests auseinander. Anschließend erörtern wir in Abschnitt II.4 den Themenkomplex der Interraterreliabilität bei der Bewertung von Schülertexten. Hier werden neben einer Klärung grundlegender Begriffe und einer Abgrenzung von Interraterreliabilität und Interraterübereinstimmung auch verschiedene Koeffizienten zur Ermittlung entsprechender Kennwerte und Aspekte von Raterschulungen thematisiert. In Abschnitt II.5 stellen wir aus der internationalen Literatur zusammengetragene Überlegungen zur holistischen sowie analytischen Kodierung vor und diskutieren spezifische Vorteile und Probleme beider Strategien. Abschnitt II.6 widmet sich der Formulierung spezifischer Fragestellungen, die in diesem Beitrag beantwortet werden sollen.

Die Daten einer Beispielaufgabe, die entsprechenden Testinstrumente, die untersuchte Stichprobe sowie die gewählten Raterdesigns und die verwendeten Analysemethoden werden in Abschnitt II.7 näher ausgeführt. Anschließend illustrieren wir in Abschnitt II.8 die Umsetzung der holistischen sowie der analytischen Kodierstrategie und stellen die originalen Kodieranweisungen für mehrere Beispielvariablen vor. In Abschnitt II.9 legen wir die für die einzelnen Variablen erzielten Werte der Interraterreliabilität und vergleichend Ergebnisse zu den alternativen Möglichkeiten der Kodierung dar. Im abschließenden Abschnitt II.10 werden diese Befunde vor fachdidaktischem und diagnostischem Hintergrund diskutiert und Ausblicke auf künftige Forschungsvorhaben eröffnet.

II.2 Schreibkompetenz – Konstrukt und Erwerb

Das Schreiben zählt in unserer Gesellschaft zu den grundlegenden Kulturtechniken und ist ein bedeutendes Mittel der Kommunikation zwischen Menschen. Es befähigt uns dazu, Wissensbestände, Gedanken, Gefühle, Erlebnisse und Erfahrungen aufzuzeichnen und sie auf diese Weise einer Wiederholung, einer späteren Betrachtung oder einer nachträglichen Reflexion zugänglich zu machen. In einer demokratischen Gesellschaft stellt die Schreibkompetenz darüber hinaus eine wichtige Möglichkeit dar, seine Meinung zu äußern, diese argumentativ überzeugend darzulegen und als mündiger Bürger mit anderen in Diskurs zu treten (vgl. U.S. Department of Education, 2003).

Schreibkompetenz ist eine hochkomplexe Fähigkeit, die sich aus verschiedenen Teilkompetenzen zusammensetzt (vgl. Becker-Mrotzek & Böttcher, 2006; vgl. Steffen, 1995; vgl. Weigle, 2002). So verstehen beispielsweise Becker-Mrotzek und Böttcher (2006) Schreibkompetenz als Ergebnis des Zusammenwirkens verschiedener Komponenten, wie grammatischer und lexikalischer Kenntnisse, Textmuster- und Schriftkenntnisse sowie sozialer Kognition. Im Rahmen der Dokumentation der DESI-Studie definieren Harsch, Neumann, Lehmann und Schröder (2007) Schreibkompetenz als die Fähigkeit, „Texte adressatengerecht zu formulieren und, je nach Zielsetzung, präzise zu informieren, überzeugend zu argumentieren oder Sprache ästhetisch ansprechend und kreativ einzusetzen“ (Harsch et al., 2007, S. 53), und betonen somit die Vielschichtigkeit von Schreibanlässen und Textsorten.

In den letzten Jahrzehnten sind verschiedene Modelle der Entwicklung, der „Stufen“ beziehungsweise „Niveaus“ sowie der Dimensionen, also der Facetten der Schreibkompetenz im Allgemeinen beziehungsweise im Hinblick auf einzelne Textsorten entwickelt worden (vgl. Augst & Faigel, 1986; Bachmann, 2002; Becker-Mrotzek, 1997; Bereiter, 1980; Bereiter & Scardamalia, 1987; Feilke, 1996, 2003; Feilke & Schmidlin, 2005; Fix, 2000; Hug, 2001; Jechle, 1992; Schneuwly, 1988). Hinzu kommen Modelle des Schreibprozesses, unter anderem auf der Basis des Novizen- und Experten-Paradigmas (vgl. Sieber, 2003).

Das international bekannteste und am intensivsten diskutierte Entwicklungsmodell der Schreibkompetenz, welches den kognitiven Aspekt der Entwicklung fokussiert, stammt von Carl Bereiter (1980). Dieser sieht sich in der Tradition Piagets. Er fragt, welche kognitiven Strategien beim Schreiben im Verlauf der Entwicklung unter der Voraussetzung zum Zuge kommen, dass die Informationsverarbeitungskapazität jeweils begrenzt ist. Mit wachsendem Alter sind wir fähig, eine größere Zahl von Operationen zu koordinieren. Das setzt – so die Annahme – voraus, dass Tätigkeiten untergeordneter Art automatisiert werden können beziehungsweise wenigstens schrittweise geringerer Aufmerksamkeit bedürfen. Nach Bereiters Ansicht sind beim entwickelten Schreiben von Experten sechs Teilfähigkeitskomplexe integriert:

- Flüssigkeit der Produktion geschriebener Sprache,
- Flüssigkeit im Bereitstellen von Wissen,
- Beherrschen von Schreibkonventionen,
- Übernahme der Perspektive von anderen (d. h. Lesern),
- Bewertung von Texten sowie
- metakognitives Denken.

Die Fähigkeiten, flüssig zu schreiben und Wissen bereitzustellen, konstituieren die basale „Strategie" des assoziativen Schreibens (*Associative Writing*). Kennzeichnend ist, dass man schreibt, was einem spontan in den Sinn kommt. Dies schließt auch Wiederholungen ein. Auf Aspekte wie Kohäsion und Kohärenz kann der junge Schreiber kaum achten, da die Aufmerksamkeit durch die Orientierung am bloßen Fortgang des Schreibens absorbiert wird. Es kommt somit allenfalls zu lokaler Kohärenzbildung.

Wird diese „inhaltszentrierte" Strategie beherrscht, setzt eine allmähliche Integration eines weiteren Fähigkeitskomplexes ein. Man orientiert sich verstärkt an Schreibkonventionen und dies nicht nur im Hinblick auf Syntax einschließlich Orthografie, sondern auch in Rücksicht auf Textsortennormen – das normorientierte Schreiben setzt ein (*Performative Writing*). In dieser Phase ruft man beispielsweise in der Schule gelehrte Textsortenmuster (z. B. Brief) mehr oder weniger rigide ab, ohne aufgabenspezifische Modifikationen bedenken zu können.

Über „Inhalt" und „Form" hinaus wird in einer nächsten Phase auch der Aspekt des potenziellen Lesers systematisch berücksichtigt. Mit der Übernahme seiner Perspektive geht die Erkenntnis einher, dass die eigene Perspektive „dezentriert" werden kann und insofern „relativ" ist. Das Schreiben wird kommunikativ (*Communicative Writing*).

Die Fähigkeit zur Dezentrierung ist Voraussetzung für das *Unified Writing*, welches von Baurmann (1995) als „authentisches", von Bachmann (2002) als „kritisches" Schreiben etikettiert wird. In diesem Entwicklungsstadium hat der Schreibende genügend Ressourcen für eine (selbst-)kritische Prüfung der Prozesse und bisherigen (Vor-)Produkte zur Verfügung, es entwickelt sich ein eigener Schreibstil. „Der Text wird nun als etwas empfunden, das zu gestalten ist, d. h. Schreiben wird nicht mehr nur als instrumentelle Fähigkeit, etwas mitzuteilen, gesehen, sondern weit stärker als eine produktive Tätigkeit" (Eigler, Merziger & Winter, 1990, S. 17).

Steht in der Phase des Unified Writing das Produkt im Fokus, so ist in einer im Modell vorgesehenen weiteren Phase wieder, wie schon beim assoziativen Schreiben, die Prozesskategorie involviert. Das Schreiben wird nicht mehr als nachträgliche Fixierung von bereits Gewusstem begriffen, sondern der Schreibende gelangt durch das Schreiben zu neuem,

„tieferem" Verstehen von Sachverhalten, das Schreiben verhilft zu neuer Erkenntnis. Diese Phase wird als epistemisches Schreiben (*Epistemic Writing*) bezeichnet.

Im Hinblick auf Bereiters Modell ist verschiedentlich Kritik geübt worden. So wurde unter anderem diskutiert, dass keine Theorie zugrunde liege, auf deren Basis plausibel werde, warum für dieses oder jenes Niveau einmal ein Prozess- und einmal ein Produktaspekt leitend sein sollen. Außerdem werde stufen- beziehungsweise niveauintern zu wenig differenziert, was insbesondere beim normorientierten Schreiben auffällt (vgl. Feilke, 2003): Gibt es nicht verschiedene „Grade", in denen man Normen gerecht zu werden vermag? Hinzu kommt, dass eine Behauptung wie die, eine bestimmte Art des Schreibens sei erkenntnisbildend, aus methodischen Gründen nur schwer zu belegen ist. Ebenso ist schwer zu entscheiden, ob ein in einem spezifischen Kontext als angemessen erscheinender „Verstoß" gegen eine Textsortennorm nun intendiert war oder nicht, ob es sich also um Unified Writing (bzw. authentisches/kritisches Schreiben) handelt oder nicht.

Andererseits kann im Rahmen dieses Modells verständlich werden, „wie durch Automatisierung und Routinisierung von Tätigkeiten auf unteren Ebenen komplexere Fähigkeiten […] ins Spiel kommen können und so die jeweils begrenzte Verarbeitungskapazität ausgelastet, aber nicht überlastet wird" (Eigler et al., 1990, S. 18f.).

II.3 Schreibkompetenz in den Bildungsstandards und Möglichkeiten ihrer Testung

In den von der KMK für das Fach Deutsch im Primarbereich verabschiedeten Bildungsstandards wird die von den Grundschülerinnen und Grundschülern der vierten Jahrgangsstufe erwartete Schreibkompetenz wie folgt erläutert: „Sie [die Schülerinnen und Schüler; Anmerkung v. Verf.] gestalten den Schreibprozess selbstständig und verfassen ihre Texte bewusst im Zusammenhang von Schreibabsicht, Inhaltsbezug und Verwendungszusammenhang" (KMK, 2005, S. 8). Konkret werden unter dem Stichwort „Schreiben" die Teilkompetenzen „über Schreibfertigkeiten verfügen", „richtig schreiben" und „Texte verfassen" benannt.

In diesem Beitrag steht der dritte Teilaspekt im Vordergrund, welcher die Komponenten Planen, Verfassen und Überarbeiten von Texten als Substandards umfasst. Das schulisch-curricular intendierte Schreiben wird in der Regel als komplexer Problemlöseprozess begriffen, wobei die genannten Teilprozesse als iterativ-rekursiv verstanden werden. Eine detaillierte Aufschlüsselung der jeweiligen Charakterisierung der Subkompetenzen kann der Tabelle II.1 entnommen werden.

II.3.1 Messung von Schreibkompetenzen

Derzeit fokussiert die standardbasierte Diagnostik der Schreibkompetenz im Primarbereich auf das Schreibprodukt, nicht jedoch auf den Schreibprozess, was in erster Linie durch die Limitationen der Testsituation zu erklären ist. Mittelfristig ist es aber wünschenswert, auch die Prozessaspekte stärker in die Diagnostik zu integrieren. Um aber auch gegenwärtig große Teile der in den Standards thematisierten Teilkompetenzen überprüfbar zu machen, wurden in den Schreibaufgaben verschiedene Schreibanlässe ebenso wie verschiedene Adressaten berücksichtigt. Auf diese Weise gestatten die verfügbaren Schreibaufgaben Aussagen über die Kompetenz der Kinder, „Erlebtes und Erfundenes; Gedanken und Gefühle; Bitten, Wünsche, Aufforderungen und Vereinbarungen; Erfahrungen und Sachverhalte“ niederzuschreiben, sowie über die Fähigkeit, „nach Anregungen (Texte, Bilder, Musik) eigene Texte“ zu verfassen (KMK, 2005, S. 11).

Tabelle II.1: Substandards des Bereichs „Texte verfassen“ und ihre Charakterisierung (vgl. KMK, 2005, S. 11)

Substandard	Charakterisierung
Texte planen	Schreibabsicht, Schreibsituation, Adressaten und Verwendungszusammenhang klären
	sprachliche und gestalterische Mittel und Ideen sammeln: Wörter und Wortfelder, Formulierungen und Textmodelle
Texte schreiben	verständlich, strukturiert, adressaten- und funktionsgerecht schreiben: Erlebtes und Erfundenes; Gedanken und Gefühle; Bitten, Wünsche, Aufforderungen und Vereinbarungen; Erfahrungen und Sachverhalte
	Lernergebnisse geordnet festhalten und auch für eine Veröffentlichung verwenden
	nach Anregungen (Texte, Bilder, Musik) eigene Texte schreiben
Texte überarbeiten	Texte an der Schreibaufgabe überprüfen
	Texte auf Verständlichkeit und Wirkung überprüfen
	Texte in Bezug auf die äußere und sprachliche Gestaltung und auf die sprachliche Richtigkeit hin optimieren
	Texte für die Veröffentlichung aufbereiten und dabei auch die Schrift gestalten

Intuitiv scheint eine Erfassung der Schreibkompetenz nur über den Einsatz freier Aufgabenformate möglich. Aus methodischer Sicht ist aber auch eine indirekte Operationalisierung dieses Konstrukts als Alternative zum Verfassen von Aufsätzen denkbar. Hierbei kommen geschlossene Itemformate oder Kurzantworten zum Einsatz, um den Stand des explizierbaren Wissens in Bezug auf Schreibkonventionen oder den Schreibprozess zu ermitteln (vgl. Weigle, 2002). Dieses Vorgehen ermöglicht zwar die Nutzung der generellen Vorteile des geschlossenen Itemformats, wie beispielsweise die sehr hohe Auswertungsobjektivität, stößt im Grundschulbereich jedoch schnell an seine Grenzen, da das explizierbare schreibbezogene Wissen der Schülerinnen und Schüler zu diesem Zeitpunkt ihrer Schullaufbahn noch sehr begrenzt ist. Entscheidender ist aber, dass bei dieser Form der Operationalisierung ein verändertes Konstruktverständnis zugrunde liegt, welches ausschließlich bestimmte Wissensaspekte der Schreibkompetenz thematisiert. Zudem erzeugt diese indirekte Art der Schreibkompetenzdiagnostik oftmals ausgeprägte Akzeptanzprobleme bei Lehrkräften und fachdidaktischen Experten.

Die direkte Erfassung der Schreibkompetenz mithilfe von Schüleraufsätzen birgt allerdings die komplexe Herausforderung der Auswertung und Beurteilung der geschriebenen Texte durch Beurteiler (Rater)[1]. Diese Problematik gestaltet sich umso komplizierter, je offener die Aufgabenstellung formuliert ist und je variabler daher die Schülerantworten sind (vgl. Mullis, Martin & Kennedy, 2004).

II.4 Das Problem der reliablen Beurteilung von Schüleraufsätzen

Die Bewertung von Schülertexten birgt die Gefahr, dass sie maßgeblich durch den subjektiven Eindruck bestimmt wird, den der Rater vom Text gewonnen hat. Dies bedeutet, dass eine Beurteilung freier Antworten im Allgemeinen und insbesondere die Beurteilung von Aufsätzen im ungünstigsten Fall eher die Meinung des Raters und weniger die eigentlich ausschlaggebende Schülerfähigkeit widerspiegelt. Werden zum Beispiel die Texte zweier Schüler vergleichbarer Fähigkeit von verschiedenen Ratern beurteilt, die verschiedene Maßstäbe anlegen, können trotz gleicher Kompetenz der Schüler unterschiedliche Fähigkeitsbeurteilungen resultieren. Dies ist aus messtheoretischen Gründen sowie im Hinblick auf die Gewährleistung einer fairen Testung äußerst problematisch.

[1] In diesem Beitrag werden Rater als Personen verstanden, die Bewertungen von Schreibprodukten vornehmen. Die Bezeichnungen Rater und Beurteiler werden im Folgenden synonym gebraucht.

In der Literatur werden zahlreiche Beurteilungsfehler (z. B. Tendenz zur Strenge bzw. Milde, Tendenz zur Mitte, Tendenz zur konsistenten Bewertung (Halo-Effekt), logische Fehler beim Erinnern, Wishful Thinking, Verfügbarkeitsfehler etc.) diskutiert, die hier nicht ausführlich erörtert werden können (für einen Überblick vgl. Kahneman, Slovic & Tversky, 1982; vgl. Myers, 2002). Allein der Halo-Effekt (vgl. Thorndike, 1920) soll kurz näher beleuchtet werden. Hierbei erzeugen einzelne Merkmale des Schülertextes einen bestimmten Gesamteindruck, der die Wahrnehmung und Beurteilung anderer Teilaspekte des Textes überstrahlt. Den Ratern gelingt es also nicht, zwischen konzeptuell verschiedenen Teilaspekten zu differenzieren. Vielmehr beurteilen sie die Schülerantwort aufgrund eines gewonnenen Gesamteindrucks, der sich aus einzelnen, besonders hervorstechenden oder als herausragend relevant eingeschätzten Besonderheiten speist. Für verschiedene Teilleistungen, die eigentlich getrennt voneinander bewertet werden sollen, vergeben die Rater dann sehr ähnliche Bewertungen (vgl. Engelhard, 1994; vgl. Knoch, Read & von Randow, 2007). Nimmt man keine Korrektur des Halo-Effektes vor, so ist von einer Überschätzung der Korrelationen zwischen Variablen auszugehen. Außerdem ergeben sich dadurch oftmals fälschlicherweise erhöhte Reliabilitätsschätzungen.

Wirtz und Caspar (2002) nennen im Wesentlichen zwei Ursachen für abweichende Urteile. Die erste mögliche Ursache bezieht sich auf die operationale Definition des einzuschätzenden Konstrukts. Da es sich bei einem Schülertext als Indikator der Schreibkompetenz um ein Produkt aus einer Vielzahl von kognitiven Teiloperationen handelt, gilt es, „eine große Zahl an Leistungsaspekten zu identifizieren und zu bewerten“ (Grzesik & Fischer, 1984, S. 77). Rater können sich nun zum einen darin unterscheiden, welche Aspekte sie berücksichtigen, und zum anderen können sie voneinander im Hinblick auf die vorgenommene Gewichtung dieser Aspekte abweichen. Als ein Grund lässt sich also festhalten, dass Rater in der operationalen Definition des Konstrukts nicht übereinstimmen.

Als zweite mögliche Ursache mangelnder Übereinstimmung wird eine unterschiedliche Einschätzung der Ausprägung des zu beurteilenden Merkmals diskutiert. Hier besteht also kein Dissens im Hinblick darauf, was Bestandteil des Konstrukts ist und wie diese Einzelbestandteile zu gewichten sind, vielmehr herrscht Uneinigkeit in Bezug auf das Ausmaß gezeigter Kompetenz.

Im Sinne der Gütekriterien psychometrischer Tests ist das Ziel eine ausreichende Reliabilität der Beurteilungen. Der Begriff der *Interraterreliabilität* bezeichnet hierbei die Übereinstimmung von Urteilen zwischen verschiedenen Ratern. Eine Beurteilung ist dann reliabel, wenn verschiedene Rater mit gleichem Wissensstand im Hinblick auf ein und denselben Schülertext zu einem ähnlichen Urteil kommen. Ausschlaggebend dürfen also nicht die unterschiedlichen Meinungen der Rater, sondern ausschließlich die Ausprägungen der Schreibkompetenz der zu beurteilenden Schüler sein. Oftmals wird in diesem Zusammenhang die

Forderung aufgestellt, dass Rater untereinander austauschbar sein sollen. Eine solche Austauschbarkeit ist dann gegeben, wenn die Unterschiede zwischen den Urteilen verschiedener Rater für dieselbe Schülerantwort vernachlässigbar klein sind.

In der Literatur wird zuweilen zwischen Beurteiler- oder Interraterreliabilität (IRR) und Beurteilerübereinstimmung oder Interrater Agreement (IRA) unterschieden (vgl. LeBreton & Senter, 2008; vgl. Wirtz & Caspar 2002). Hierbei bezieht sich die Interraterreliabilität auf die Frage, inwieweit verschiedene Beurteiler eine relative Konsistenz in ihren Urteilen zeigen und zu einer ähnlichen Rangreihe in Bezug auf die Fähigkeitseinschätzung der getesteten Schülerinnen und Schüler kommen (Relative Consistency). Bei der Beurteilerübereinstimmung dagegen steht zur Debatte, inwieweit Beurteiler absolut übereinstimmen und damit tatsächlich austauschbar sind (Absolute Consensus). Insbesondere beim Einsatz von mehrstufigen Ratingskalen mit beispielsweise fünf oder sechs Stufen, auf denen die Güte einer erbrachten Leistung eingeschätzt wird, ist eine exakte Übereinstimmung äußerst selten. Die Forderung nach einer exakten Übereinstimmung ist in einem solchen Fall aber weder notwendig noch sinnvoll. Erforderlich ist vielmehr, dass alle Rater durch ihre Urteile dieselbe Anordnung der Aufsätze in einer Rangreihe erzeugen und sich hinsichtlich der relativen Qualität der Schülerantworten einig sind. Mitunter wird gefordert, dass sich die Rater bei der Bewertung der Fähigkeitsausprägung um nicht mehr als eine Kategorie der Skala unterscheiden sollte, sodass die Kompetenzen zwar nicht identisch, aber sehr ähnlich eingeschätzt werden.

Wichtig ist auch, dass ein Rater über alle von ihm eingeschätzten Schülertexte beziehungsweise über verschiedene Messzeitpunkte hinweg konsistent urteilt. Dies bedeutet, dass er seine operationale Definition des Konstrukts sowie seine Maßstäbe im Hinblick auf die Ausprägung des Merkmals im zeitlichen Verlauf der Ratingprozedur keinesfalls ändern darf, sondern stabil halten muss. Dieses Problem wird in der Literatur unter dem Stichwort der *Intraraterreliabilität* diskutiert (vgl. Congdon & McQueen, 2000).

Um Beurteilungsfehlern und Ratereffekten entgegenzuwirken, ist es nötig, ein kriterial orientiertes Beurteilungsverfahren zu entwickeln, welches den Einfluss individueller Verzerrungen minimiert (vgl. hierzu auch Abschnitt II.5). Weiterhin bedarf es in jedem Fall der Durchführung von Mehrfachkodierungen (vgl. Lumley & McNamara, 1995), die in einem adäquaten Raterdesign (vgl. Hoyt, 2000) arrangiert werden müssen. Ferner müssen die Rater intensiv geschult und mit den Aufgaben, den theoretischen Hintergründen sowie insbesondere mit den Kodieranweisungen vertraut gemacht werden. Raterschulungen dienen somit der Einführung in die operationale Definition des Konstrukts und der Etablierung eines gemeinsamen Maßstabs in Bezug auf die Ausprägung des Merkmals. Durch Probekodierungen zwischen wiederholten Schulungsterminen lassen sich besonders kontroverse Schülertexte

identifizieren. Solche Texte beim nächsten Schulungstermin in der Gruppe zu diskutieren hat sich unserer Erfahrung nach als besonders gewinnbringend erwiesen. Es wird aber auch berichtet, dass Raterschulungen nur in beschränktem Maß dazu geeignet sind, Variabilität zwischen den Urteilen verschiedener Rater deutlich zu reduzieren oder sogar eine Austauschbarkeit der Rater herzustellen (vgl. Barret, 2001; vgl. Lumley & McNamara, 1995; vgl. Weigle, 1998). Insgesamt gelingt es durch Schulungen eher, die raterinterne Konsistenz zu erhöhen und das Auftreten von Ausreißern zu minimieren, als für alle Schülertexte stets eine absolute Übereinstimung aller Rater zu erzielen (vgl. Weigle, 1998).

II.4.1 Maße der Interraterreliabilität und Beurteilerübereinstimmung

Die Übereinstimmung beziehungsweise Reliabilität zwischen den Ratern kann mithilfe verschiedener Maße quantifiziert werden. Für die Ermittlung der Beurteilerübereinstimmung werden in Abhängigkeit vom Datenniveau häufig die absolute und prozentuale Übereinstimmung, die Übereinstimmung bei der Wahl benachbarter Kategorien, Cohens κ (Kappa), adjustierte κ-Koeffizienten sowie Kendalls τ (Tau) verwendet. Als Maß der Interraterreliabilität wird oft die Intraklassenkorrelation (*Intra Class Correlation*, ICC) berichtet.

Die prozentuale Übereinstimmung ist das einfachste Übereinstimmungsmaß und beziffert den prozentualen Anteil der Fälle, in denen zwei Rater ein identisches Urteil abgeben (vgl. Fleiss, 1973). Sie eignet sich für kategoriale Daten und wird berechnet als Quotient aus der Anzahl der übereinstimmend bewerteten Schülertexte und der Anzahl aller bewerteten Schülertexte, multipliziert mit 100 Prozent. Im Zusammenhang mit der prozentualen Übereinstimmung erscheint problematisch, dass sie in hohem Maße dadurch beeinflusst wird, welche Grundwahrscheinlichkeit ein einzuschätzendes Merkmal besitzt (vgl. Hayes & Hatch, 1999). Wird beispielsweise ein bestimmter dichotom einzuschätzender Teilaspekt der Schreibkompetenz bereits von sehr vielen Schülerinnen und Schülern sehr gut beherrscht, dann werden die beiden Kategorien „1“ (= wird beherrscht) und „0“ (= wird nicht beherrscht) nicht mit der gleichen Wahrscheinlichkeit von je 50 Prozent von einem Rater gewählt. Vielmehr ergeben sich stark verzerrte Kategorienbelegungen, sodass die Kategorie „1“ deutlich stärker besetzt ist als die Kategorie „0“. Ein solches Missverhältnis kann die prozentuale Übereinstimmung nicht abbilden. Vielmehr liegt bereits die durch Zufall erwartbare prozentuale Übereinstimmung sehr hoch, ohne dass sich die Rater tatsächlich einig sein müssen. Hier sollte nun die erreichte Übereinstimmung zweier Rater an der durch Zufall erreichbaren Übereinstimmung relativiert werden. Ein solches zufallsbereinigtes Maß der Raterübereinstimmung ist Cohens κ. Es berücksichtigt das Verhältnis der beobachteten zu der durch Zufall erreichbaren Übereinstimmung. Der Koeffizient Cohens κ

ist ein standardisiertes Maß, welches Werte zwischen –1 und +1 annehmen kann und für nominale Daten geeignet ist. Als Maß für Beurteilerübereinstimmung findet Cohens κ sehr häufig Anwendung, wird aber auch kritisch diskutiert (vgl. Übersax, 2008).

Die Intraklassenkorrelation schließlich basiert auf einer Zerlegung von Varianzen und eignet sich für intervallskalierte Daten. Bei der Zerlegung der Varianzkomponenten (vgl. Brennan, 2001) können im Wesentlichen drei mögliche Quellen unterschieden werden. Dies sind:

- die Varianz der Schülerfähigkeiten, die sich in den Schülertexten niederschlägt;
- die Varianz aufseiten der Rater, beispielsweise Unterschiede im Hinblick auf Strenge oder Milde der Beurteilung, und
- die Varianz der Dyade, also Unterschiede in der Wechselwirkung zwischen Rater und Schülertext (vgl. Hoyt, 2000).

Die ICC gibt nun den Teil der um die Raterstrenge bereinigten Gesamtvarianz an, der auf die wahre Varianz in der Schreibkompetenz der Schülerinnen und Schüler zurückführbar ist. In der einschlägigen Literatur wird dieser Varianzanteil auch als Varianz des *Latent Trait* bezeichnet. Die ICC kann Werte zwischen 0 und +1 annehmen und lässt sich als mittlere Korrelation zwischen Ratern interpretieren.

Für die genannten Maße finden sich in der Literatur verschiedene Empfehlungen, welche Ausprägung als Mindestanforderungen und welche Bereiche als zufriedenstellende oder sogar sehr gute Beurteilerübereinstimmung interpretiert werden können. Hierbei handelt es sich aber nicht um verbindliche Grenzwerte. Diese können nicht in allgemeingültiger Form festgelegt werden, da immer der jeweilige Forschungskontext sowie die spezifischen Merkmale der Studie ausschlaggebend sind. Für den hier vorliegenden Kontext der Beurteilung von Schüleraufsätzen im Rahmen von Large-Scale-Assessments berichtet Neumann (2007) für fünfstufige Skalen (wie Globalurteil, Textaufbau, Ausdruck etc.) Interraterkorrelationen, die mit der ICC vergleichbar sind. Für die Daten der Studie LAU11/ULME1[2] bewegen sich die Interraterkorrelationen bei der Bewertung von Briefen in einem Bereich von $r = .38$ bis .54 bei einer mittleren Korrelation von $r = .44$. Für ganz ähnlich konzipierte Schreibaufgaben wurden im Rahmen der DESI-Studie für den ersten Messzeitpunkt mittlere Interraterkorrelationen von $r = .66$ bis .70 mit einem Wertebereich von $r = .61$ bis .75 ermittelt. Bereits anhand dieser Beispiele zeigt sich die Stichprobenspezifität solcher Befunde, die sich in der deutlich größeren Varianz der

[2] LAU11 ist ein Akronym für die Studie „Aspekte der Lernausgangslage und Lernentwicklung“ in der elften Jahrgangsstufe. ULME1 steht für „Untersuchung der Leistungen, Motivation und Einstellungen zu Beginn der beruflichen Ausbildung“.

Schülerleistungen in DESI im Vergleich zu ULME äußert, sodass Unterschiede zwischen Schülern der Stichprobe zuverlässiger in DESI festgestellt werden konnten. Inwieweit diese Ergebnisse auch für den Grundschulbereich erwartbar sind, in dem zum einen wesentlich weniger geschrieben wird und zum anderen auch weniger Textsortennormen berücksichtigt werden, ist fraglich.

Wie bestimmte Größenordnungen an Übereinstimmung von verschiedenen Autoren interpretiert wurden, stellen Wirtz und Caspar (2002) beispielhaft für den Koeffizienten Cohens κ vor. Diese Angaben sind in Tabelle II.2 zusammengefasst.

Tabelle II.2: Richtwerte zur Beurteilung von Cohens κ nach Wirtz und Caspar (2002)

Quelle (Anwendungsbereich)	Güte der Übereinstimmung	κ-Wert bzw. κ-Bereich
Fleiss & Cohen, 1973	sehr gut	κ > .75
	gut	κ > .60
	noch akzeptabel	κ > .40
Margraf & Fehm, 1996 (für den klinischen Bereich)	gut	κ > .70
	zufriedenstellend	κ > .50
Bakeman & Gottman, 1986	zufriedenstellend	κ > .70
Frick & Semmel, 1978	zufriedenstellend	κ > .75

Es fehlen aber nicht nur klare Richtwerte für einzelne Koeffizienten, sondern auch bindende Vorgaben, wann welche Maße für die Reliabilität oder Übereinstimmung der Rater ermittelt werden sollten. Die Wahl des berichteten Koeffizienten sollte sich daher am vorliegenden Datenniveau und der jeweiligen Fragestellung orientieren, da die verschiedenen Koeffizienten unterschiedliche Aspekte in den Vordergrund stellen und ganz verschiedene Vor- und Nachteile aufweisen.

Im Zusammenhang mit der Berechnung der Beurteilerübereinstimmung oder der Interraterreliabilität möchten wir betonen, dass die ermittelte Reliabilität keine Eigenschaft des Messinstruments, also z. B. der Ratingskala selbst, sondern eine Eigenschaft ihres Einsatzes in einer konkreten Studie mit bestimmten Ratern und einer bestimmten Schülerstichprobe ist. Eine Reliabilitätsangabe für eine bestimmte Ratingskala darf somit nicht in dem Sinne missverstanden

werden, als wäre die Reliabilität eine inhärente Eigenschaft der Skala, welche unabhängig von der Fähigkeit der Rater vorliegt, die die Skala verwenden (vgl. Eckes, 2008; vgl. Overall & Magee, 1992).

Einen ausführlichen Überblick über verschiedene Maße der Beurteilerübereinstimmung und der Beurteilerreliabilität, ihre Eigenschaften und Einsatzmöglichkeiten sowie Formeln zu ihrer Berechnung findet man beispielsweise bei Übersax (2008) sowie für den deutschsprachigen Raum bei Wirtz und Caspar (2002).

II.5 Varianten der Kodierung

Allgemein können bei der Bewertung von Schüleraufsätzen grob zwei Strategien der Bewertung oder Kodierung unterschieden werden: die holistische (vgl. Cooper, 1977) und die analytische Kodierung (vgl. Bryant & Bryant, 2003; vgl. Lloyd-Jones, 1977).

Bei der holistischen Kodierung wird vom Kodierer ein Globalurteil abgegeben, welches die Leistung des Schülers insgesamt auf einer oft mehrstufigen Ratingskala einordnet. Diese Globalbeurteilung ist allerdings keine subjektive Ad-hoc-Einschätzung, die allein den ersten intuitiven Eindruck eines Raters wiedergibt. Vielmehr basiert auch ein holistisches Gesamturteil auf kriterialen Vorgaben und kann entweder mithilfe von Benchmarktexten, durch die Berücksichtigung von vorab klar definierten Kriterien oder durch eine Kombination aus beidem erfolgen. Eine solche holistische Bewertung der Ausprägung eines Merkmals mithilfe einer Ratingskala wird auch als hoch inferentes Rating bezeichnet und führt zu einer eher größeren erfassten Gesamtvarianz bei gleichzeitig geringerer – aber dennoch zufriedenstellender – Beurteilerübereinstimmung. So wurden Varianten holistischer Kodierung lange Zeit in erster Linie aufgrund ihrer guten Beurteilerreliabilitäten eingesetzt (vgl. Huot, 1990). Hiervon unberührt liegt eines der größten Probleme holistischer Kodierung in dem Umstand begründet, dass „[…] holistic scoring, which concentrates on the general impact of a text, may not be especially sensitive to particular textual features of student writing“ (Huot, 1990, S. 209). Ein holistischer Eindruck, beispielsweise der, dass eine unzureichende Schreibkompetenz vorliegt, gibt daher keinerlei Anhaltspunkte, wo eine notwendige Förderung sinnvoll ansetzen könnte.

Bei der analytischen Kodierung wird ein Konstrukt sehr fein ausdifferenziert und in seine einzelnen Bestandteile zerlegt. Diese Einzelbestandteile, wie z. B. bestimmte inhaltliche oder sprachliche Aspekte, werden dann zumeist dichotom beziehungsweise auf Skalen mit nur wenigen – beispielsweise drei – Stufen hinsichtlich ihres Auftretens bewertet und anschließend aggregiert. Beispielsweise können für die Beurteilung des Inhalts eines Aufsatzes vorgegebene Schlagwörter oder Schlüsselphrasen im Text identifiziert und diese einzeln und unabhängig

voneinander hinsichtlich ihres Vorhandenseins dichotom (0/1) kodiert werden. So verstanden sind einzelne Wörter beziehungsweise Wortgruppen die Basis einer analytischen Kodierung (vgl. Slater & Boulet, 2001). Solche wenig komplexen Kodierungen, die auch als niedrig inferente Ratings bezeichnet werden, führen zu eher wenig erfasster Gesamtvarianz, haben aber den Vorteil höherer Beurteilerübereinstimmung. Goulden (1994) berichtet allerdings vergleichbare Reliabilitäten für holistische und analytische Kodierungen, wobei sich mitunter ein gewisser Vorsprung analytischer Kodierungen abzeichnet.

Die Definition und Abgrenzung von analytischen und holistischen Urteilen gestaltet sich mitunter problematisch. So wurde beispielsweise im Kontext der Testung der Bildungsstandards für Englisch als erste Fremdsprache bei der Erhebung der Schreibkompetenz ebenfalls mit holistischen und analytischen Kodierungen gearbeitet. In dieser Studie wurden neben einem dreistufigen holistischen Gesamteindruck auch vier dreistufige, so genannte analytische Variablen zu den Aspekten Aufgabenerfüllung (*Task Fulfilment*), Organisation des Textes (*Organization*), Grammatik (*Grammar*) und Wortschatz (*Vocabulary*) erhoben. Der Aspekt der Organisation des Textes bezieht sich neben der optischen Strukturierung des Textes auf die Entwicklung des Plots beziehungsweise der Argumentation. Zusätzlich werden die Verwendung von Kohäsionsmitteln und die sprachliche Gestaltung thematisiert (vgl. Rupp, Vock, Harsch & Köller, 2008). Hier sind für die Beurteilung der Organisation des Textes also nicht alle verfügbaren Informationen relevant, so sollen zum Beispiel keine inhaltlichen Aspekte in die Bewertung einbezogen werden. Dennoch müssen für die Bewertung gleichzeitig verschiedene Teilaspekte berücksichtigt, gewichtet und zu einem Urteil integriert werden. Dies entspricht nicht unserem Verständnis einer analytischen Kodierung. In Übereinstimmung mit Slater und Boulet (2001) erachten wir einzelne Wörter oder Sätze als die Grundlage eines analytischen Kodierzugangs und nicht den Gesamttext. Dementsprechend werden Variablen, die Stil, Inhalt oder sprachliche Richtigkeit insgesamt betreffen und sich aus der Begutachtung mehrerer Einzelkriterien zusammensetzen, welche gewichtet und integriert werden müssen, in diesem Beitrag als holistisches Urteil für einen bestimmten Bereich, wie zum Beispiel „holistische Einschätzung des Inhalts", bezeichnet (vgl. den Abschnitt „Holistische Kodierung").

Unabhängig davon, welche Kodierstrategie favorisiert wird, besteht bei der Auswertung von Schülertexten immer die Notwendigkeit, alle zu erfassenden (Teil-)Kompetenzen explizit zu benennen, die zu beurteilenden Kompetenzabstufungen detailliert und eindeutig zu beschreiben und möglichst mithilfe von Benchmarktexten im Sinne von Schülerbeispieltexten oder Textauszügen zu illustrieren. Diese Verbindung von abstrahierenden Beschreibungen und idealtypischen oder Grenzfall-Beispieltexten erleichtert den Ratern die Entwicklung einer an expliziten Vorgaben orientierten Norm, wann welche Kompetenzbeurteilung zu vergeben ist. Für

eine Annäherung dieser Maßstäbe der einzelnen Rater an eine kollektive Normvorstellung sind eine gemeinsame Schulung sowie eine anschließende Trainingsphase unerlässlich, da sich auf diese Weise eine konsensuelle Vorstellung davon entwickeln kann, wie kodiert werden soll (vgl. Lumley, 2002)[3]. Diese geteilte Sichtweise ist neben einem adäquaten Kodiersystem eine wesentliche Voraussetzung für eine reliable Bewertung der Schüleraufsätze (vgl. Nitko, 2004).

II.6 Fragestellungen

Im Rahmen der dargestellten theoretischen Grundlagen und erörterten Problemstellungen sind für uns folgende Fragen zentral:

- Welche Befunde ergeben sich für die analytische sowie die holistische Kodierstrategie im Hinblick auf Beurteilerübereinstimmung und Beurteilerreliabilität?
- Gibt es unter Berücksichtigung der Reliabilität der Messungen einen klaren Vorteil der einen gegenüber der anderen Strategie?
- Welche Aspekte der analytischen Kodierung fließen in welchem Umfang in die holistische Kodierung ein?
- Lässt sich zeigen, dass mit beiden Kodierstrategien dieselbe Kompetenz gemessen wird?
- Wie stark sind die methodenspezifischen Anteile bei der Schülerbewertung?
- In welchem Umfang sind Halo-Effekte bei beiden Kodierstrategien zu beobachten?

II.7 Anlage der Untersuchung

II.7.1 Testentwicklung und Stichprobenbeschreibung

Für die Diagnostik der Schreibkompetenz im Rahmen unserer Studie wurden 15 Aufgaben entwickelt. Es sollten narrative, argumentative und informierende Texte geschrieben werden, wobei entweder ein Bild- oder ein Textimpuls beziehungsweise eine Kombination aus beidem vorgegeben wurde. Einige Aufgaben wurden nur in der dritten, andere nur in der vierten und eine Schnittmenge in beiden Jahrgangsstufen eingesetzt. In der Studie zur Pilotierung der

[3] Dennoch kann auch durch Raterschulungen meist keine absolute Übereinstimmung erzielt werden, vgl. Abschnitt II.3.

länderübergreifenden Bildungsstandards für den Primarbereich bearbeiteten 5 032 Dritt- und Viertklässler Aufgaben aus dem Kompetenzbereich Schreiben. Pro Aufgabe liegen für jede Klassenstufe zwischen 260 und 540 Schülertexte vor. Aufgrund des verwendeten Multi-Matrix-Designs bearbeitete jedes Kind mindestens zwei Schreibaufgaben, sodass eine Verlinkung der verschiedenen Textsorten (narrativ, argumentativ und informierend) ebenso wie eine Verknüpfung der Leistungen im Kompetenzbereich Schreiben mit den anderen Kompetenzbereichen (Sprechen und Zuhören, Lesen, Sprache und Sprachgebrauch untersuchen sowie Rechtschreibung) möglich war.

II.7.2 Beispielaufgabe

Für die im Folgenden näher betrachtete Beispielaufgabe „Die zwei Esel" liegen aus der Pilotierungsstudie im Jahr 2006 Texte von 260 Drittklässlern[4] vor, von denen 137 (52,7 Prozent) Mädchen waren. Bei der Beispielaufgabe handelt es sich um eine Bilderaufgabe, welche die Standards „nach Anregungen (Texte, Bilder, Musik) eigene Texte schreiben" und „verständlich, strukturiert, adressaten- und funktionsgerecht schreiben: Erlebtes und Erfundenes; Gedanken und Gefühle; Bitten, Wünsche, Aufforderungen und Vereinbarungen; Erfahrungen und Sachverhalte" (vgl. KMK, 2005, S. 11) operationalisiert.

Der verwendete Bildimpuls ist in Abbildung II.1 dargestellt. Die Aufgabe wurde primär dem narrativen Bereich zugerechnet. Für die Bearbeitung der Aufgabe standen den Schülerinnen und Schülern 20 Minuten zur Verfügung. Die Instruktion lautete: „Die Bilder erzählen eine Geschichte. Schreibe diese Geschichte für andere Kinder auf! Bedenke, dass die Kinder die Bilder beim Lesen nicht sehen können!" Um den Schülerinnen und Schülern eine Orientierung im Hinblick auf den erwarteten Textumfang zu geben, wurde in die Testhefte unter die Instruktion eine A4-Seite mit 21 leeren Zeilen eingefügt.

[4] In nachfolgenden Studien wurde die Aufgabe auch in Klassenstufe 4 vorgelegt.

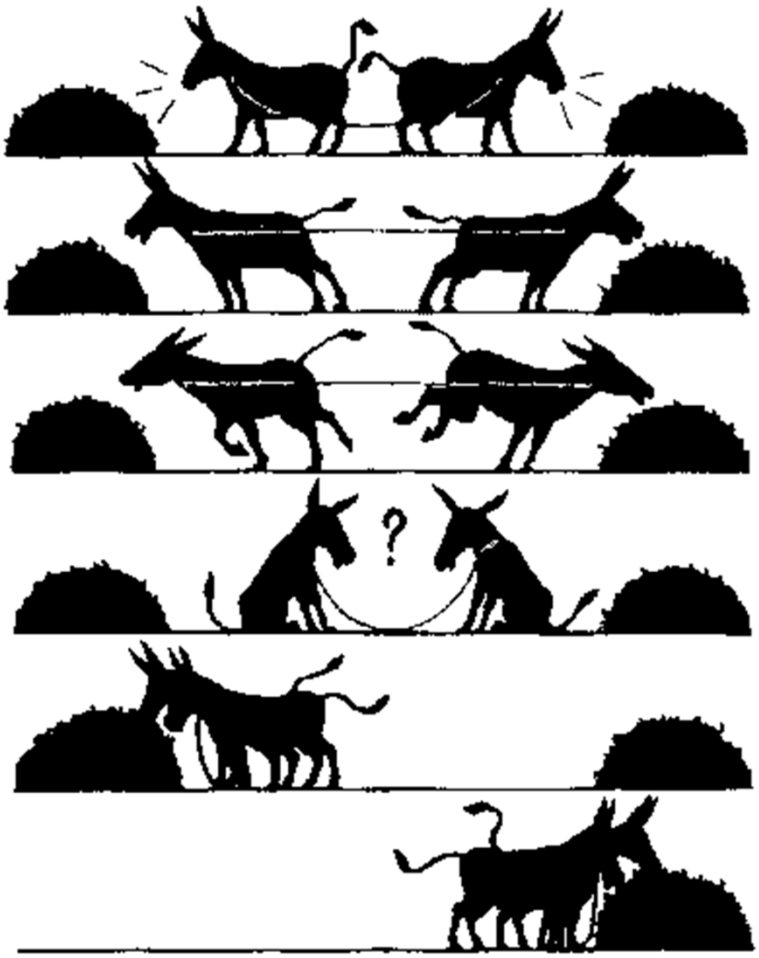

Abbildung II.1: Bildimpuls der Aufgabe „Die zwei Esel“

Die Schülerinnen und Schüler der dritten Jahrgangsstufe schrieben im Mittel 63 Wörter. Der kürzeste Text umfasste 14, der längste 143 Wörter. Lediglich in drei Prozent der Fälle, also von acht Schülerinnen beziehungsweise Schülern, wurde die Bearbeitung verweigert beziehungsweise wurde der Text erkennbar nicht zu Ende geführt. Die Abbildung II.2 zeigt eine eingescannte Originalschülerantwort.

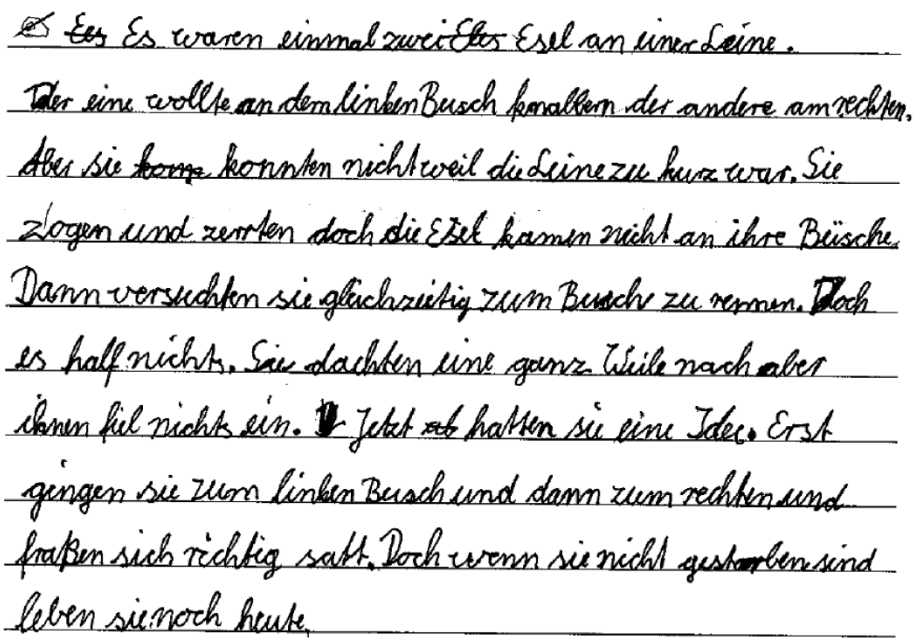
Es waren einmal zwei Esel an einer Leine.
Der eine wollte an den linken Busch knabbern der andere am rechten.
Aber sie konnten nicht weil die Leine zu kurz war. Sie
zogen und zerrten doch die Esel kamen nicht an ihre Büsche.
Dann versuchten sie gleichzeitig zum Busch zu rennen. Doch
es half nichts. Sie dachten eine ganz Weile nach aber
ihnen fiel nichts ein. Jetzt hatten sie eine Idee. Erst
gingen sie zum linken Busch und dann zum rechten und
fraßen sich richtig satt. Doch wenn sie nicht gestorben sind
leben sie noch heute.

Abbildung II.2: Originalschülerantwort zur Aufgabe „Die zwei Esel“

II.7.3 Raterdesign

Die zuverlässige Beurteilung von Schreibprodukten setzt einen Mindestanteil von Mehrfachkodierungen voraus (vgl. Lumley & McNamara 1995). Bei beispielsweise Doppelkodierungen werden für jeden Schülertext Kodierungen von zwei verschiedenen Ratern vorgenommen. Wichtig ist hierbei, dass die paarweisen Kombinationen der Rater über die Schülertexte und Variablen hinweg variiert werden. Nehmen an einer Studie beispielsweise acht Rater teil, so sind in einem ausgewogenen Design 28 Kombinationen von Ratern zu Paaren möglich. Diese denkbaren Kombinationen von Raterpaaren sollten nun gleichmäßig auf alle Schülertexte verteilt werden. Diese Strategie erlaubt hinreichend genaue Aussagen zu einer „mittleren Raterübereinstimmung“, wobei Übereinstimmung in diesem Beispiel auf eine Zufallsziehung (von acht Ratern) aus einer hypothetischen Raterpopulation bezogen ist.

Diesen Überlegungen folgend wurden in der Pilotierungsstudie alle Einzelvariablen der analytischen Kodierstrategie für jede Schülerantwort von jeweils zwei unabhängigen und intensiv geschulten Ratern bewertet. Somit liegen für jede Schülerantwort und jede Variable nicht wie sonst üblich ein, sondern zwei Angaben im Datensatz vor. Diese vollständigen Doppelkodierungen wurden nicht für alle Variablen einer Aufgabe von identischen Ratern durchgeführt, vielmehr wechseln die Kombinationen der Beurteiler so, dass eine Verlinkung der Rater untereinander über alle Variablen einer Aufgabe erfolgt. Auf diese Weise wird der

spezifische Einfluss eines bestimmten Raters, der beispielsweise besonders streng bewertet, ausgeglichen und gleichmäßig über verschiedene Variablen und verschiedene Schüler verteilt. Dadurch entstehen keine spezifischen Benachteiligungen für einzelne Schülerinnen und Schüler. Die hier beschriebenen Kodierungen der analytischen Variablen wurden von einer Gruppe von 28 Ratern des *IEA Data Processing and Research Center* (DPC) in Hamburg durchgeführt.

In einer Zusatzstudie wurde die Kodierung der analytischen Variablen mit einer distinkten Ratergruppe, der 14 Personen angehörten, am IQB wiederholt. Die Rater lieferten darüber hinaus Doppelkodierungen für die im Abschnitt „Holistische Kodierung" vorgestellten holistischen Variablen. In der IQB-Zusatzstudie wurden für die Beispielaufgabe „Die zwei Esel" ebenfalls vollständige Doppelkodierungen erhoben, nun allerdings sowohl für die analytischen als auch die holistischen Variablen. Hierbei wurde ein für Doppelkodierungen in besonderer Weise geeignetes Raterdesign, das so genannte unvollständige Blockdesign (*Incomplete Block Design*, vgl. Hoyt, 2000), verwendet, welches in Tabelle II.3 illustriert wird.

Tabelle II.3: Verlinkung der Rater innerhalb eines unvollständigen Blockdesigns für Doppelkodierungen mit vier Ratern (vgl. Hoyt, 2000)

	Rater 1	Rater 2	Rater 3	Rater 4
Text 1	X	X		
Text 2	X		X	
Text 3	X			X
Text 4		X	X	
Text 5		X		X
Text 6			X	X

II.7.4 Methodisches Vorgehen

Hinsichtlich potenzieller Rater-Effekte interessiert weniger das Verhalten einzelner Rater, sondern eher das allgemeine Funktionieren des Kodierens über alle Rater hinweg. In der Datenaufbereitung wurde daher für Übereinstimmungsanalysen und für Strukturgleichungsmodelle das Konzept der *Pseudorater* eingeführt. Dies soll zunächst erläutert werden. Arbeitet man in einer Studie beispielsweise mit acht Ratern, so sind für diese acht Beurteiler wie oben dargestellt 28 mögliche Kombinationen an Raterpaaren denkbar. Diese

Kombinationen beinhalten z. B. die Paarungen der Rater 1 und 2, 1 und 3, 1 und 4, aber auch 2 und 3, 2 und 4 etc. Betrachtet man nun alle Variablen, die in Bezug auf ein und denselben Schülertext kodiert wurden, so liegen für unterschiedliche Variablen Bewertungen aus verschiedenen Raterkombinationen vor.

Ebenso wurden nicht alle Schülertexte im Hinblick auf ein und dieselbe Variable von einem konstanten Duo, sondern von alternierenden Raterpaaren kodiert. In der Datenanalyse berücksichtigt man nun nicht länger die konkreten Kombinationen der Rater, sondern setzt für die zwei tatsächlich existenten Rater, die ja wechseln, die zwei Pseudorater ein. Pseudorater 1 steht nun immer für den „ersten" Rater eines tatsächlich existierenden Raterpaares und damit für ganz verschiedene reale Rater. Analog verkörpert Pseudorater 2 alle „zweiten" Rater, die für einen Schülertext Doppelkodierungen vorgenommen haben.

Diese Pseudorater mit korrespondierenden Pseudoratings werden als Ziehungen aus einer Menge möglicher Ratings (und damit der Rater) aufgefasst. Damit geht die Annahme der Austauschbarkeit von Ratern (vgl. Gelman, Carlin, Stern & Rubin, 2003) einher, die die Pseudorater als strukturgleich (d. h. mit gleichen Eigenschaften im Hinblick auf Verteilungsmerkmale der Variablen zur Beurteilung der Schreibkompetenz) auffasst. Deshalb werden beispielsweise Mittelwert und Varianz der beiden Pseudorater für die betrachteten Variablen als identisch modelliert. Bei diesem Vorgehen werden demnach raterspezifische Strenge (unterschiedliche Mittelwerte) und Präzision (unterschiedliche Varianzen) ignoriert und gehen in den Messfehler ein. Berechnet man die Übereinstimmung der beiden Pseudorater auf einer Variablen, so kann diese als mittlere Übereinstimmung der in der Raterstichprobe auftretenden Rater aufgefasst werden.

Im Gegensatz zu diesem Vorgehen werden in Multifacetten-Rasch-Modellen (vgl. Eckes, 2005) – grob gesprochen – verschiedene Raterstrengen (d. h. „Mittelwerte" bestimmter Variablen) einer Raterstichprobe in Rechnung gestellt. Für den Einsatz der Kodieranweisung in künftigen Assessments mit anderen Ratergruppen sind diese konkreten Raterstrengen im Vergleich zur Quantifizierung erwartbarer Messfehlervarianzen jedoch von geringerer Bedeutung. Vielmehr ist die Generalisierbarkeit der erwartbaren Variabilität eines Instrumentes bei anderen Raterpopulationen von primärem Interesse, weshalb im Falle von *Low Stakes Assessments* der Einsatz von Pseudoratern als technisches Hilfsmittel in Analysen durchaus sinnvoll ist.

Bleibt die Raterstichprobe jedoch (relativ) konstant, so ist es bei individuell relevanten Entscheidungen in *High Stakes Assessments* (z. B. Sprachtests bei TestDaF, vgl. Eckes, 2005) durchaus sinnvoll, Raterstrengen nach erfolgter Kalibrierung für individuelle Urteile in Rechnung zu stellen und damit die Rater als festen Effekt aufzufassen. Dies erlaubt die Schätzung einer um Ratereffekte bereinigten Fähigkeitsschätzung für einen Testteilnehmer. Einem Schüler mit

tendenziell strengeren Ratern werden beispielsweise etwas bessere Bewertungen im Rahmen der statistischen Analyse zugesprochen.

Die Pseudoratings auf den entsprechenden Variablen wurden zur dimensionalen Prüfung einer Hauptkomponentenanalyse (PCA, *Principal Component Analysis*) verwendet, wobei die Methode der Parallelanalyse (vgl. Bortz, 2005) zur Feststellung der Dimensionalität eingesetzt wurde. Da unter anderem dichotome Variablen vorliegen, bildeten tetrachorische Korrelationen die Basis der PCA.

Die Berechnung von Varianzkomponentenmodellen sowie von Intraklassenkorrelationen erfolgte mithilfe eines *Linear Mixed Effects Models* unter Verwendung des Pakets lme4 (vgl. Bates, Maechler & Dai, 2008) in der Software R (vgl. R Development Core Team, 2008). Für eine abhängige Variable mit mehreren Kodierungen können dabei die Faktoren Rater (hier *nicht* Pseudorater) und Schüler als zufällig spezifiziert und damit die Varianzanteile Rater, Schüler und Residuum bestimmt werden. Diese Zweiweg-Varianzanalyse wird in der Terminologie der Mehrebenenmodelle als *Cross-Classified Model* (vgl. Goldstein, 2003) bezeichnet. Der Einsatz dieser *Mixed Models* erlaubt die flexible Aufnahme weiterer Variationsquellen wie wiederholte Beurteilungen eines Raters oder die teilweise Variation von Bestandteilen einer Kodieranweisung (etwa von Benchmarks), um Einflussquellen für die Variation von Kodierungen im Rahmen einer Generalisierbarkeitstheorie (vgl. Brennan, 2001; vgl. Schoonen, 2005) abzuschätzen.

Zur Bestimmung von Korrelationen bei – aufgrund der Ratingprozesse – messfehlerbehafteten Variablen setzten wir Strukturgleichungsmodelle ein, die mithilfe des R-Pakets „sem“ (vgl. Fox, 2008) berechnet wurden. Für die Strukturgleichungsmodelle wurde dabei wiederum auf das Konzept der (austauschbaren) Pseudorater zurückgegriffen. Da diese Pseudorater strukturell gleich funktionieren, wurden Mittelwerte, Varianzen und Ladungen durch gemeinsame Parameter für beide Pseudorater beschrieben. Aufgrund dieser Modellierung kann man die mittlere Reliabilität der Raterstichprobe bestimmen. Im Sinne eines konfirmatorischen Vorgehens führten wir Analysen durch, welche die beobachtete Kovarianzstruktur zwischen Variablen und Ratern in eine Kovarianzstruktur zwischen Schülern (Between) und eine Kovarianzstruktur innerhalb von Schülern (Within) zerlegen, die die Struktur mehrfacher Ratings bei einem Schüler beschreibt.

Auf der Between-Ebene wird dabei die um raterspezifische Messfehler bereinigte „wahre Korrelation“ zwischen Variablen dargestellt. Zusammenhänge auf der Within-Ebene beschreiben dagegen Halo-Effekte (vgl. Bechger, Maris & Hsiao, 2007). Diese können als Tendenz aufgefasst werden, dass Rater, die Schüler auf einer Variablen nach Kontrolle der „wahren Schülerfähigkeit“ besser bewerten, auch auf einer anderen Variablen besser bewerten – und zwar unabhängig von der „tatsächlichen“ auf der Between-Ebene modellierten Fähigkeit. Geringe Varianzen auf der

Within-Ebene entsprechen einem geringen Messfehler und bei konstant gehaltener Fähigkeitsvarianz auf der Between-Ebene einer hohen Reliabilität einer Variablen. Neben einer möglichst hohen Reliabilität sind jedoch auch niedrige Korrelationen zwischen Variablen auf der Within-Ebene gewünscht, damit die Beurteilungen von Schülerfähigkeiten nicht durch Halo-Effekte verfälscht werden.

II.8 Umsetzung der alternativen Kodierungsvarianten

II.8.1 Analytische Kodierung

Der von uns eingesetzten analytischen Kodierung liegt die Idee zugrunde, dass Bewertungsaspekte in unterschiedlichem Maße über Textsorten hinweg verallgemeinerbar sind. Daher wird zwischen textsortenübergreifenden, textsortenspezifischen und aufgabenspezifischen Variablen unterschieden. Bei den textsortenübergreifenden Aspekten handelt es sich beispielsweise um allgemeine Sprachvariablen wie die Vielfalt der Satzkonstruktion oder die Markierung von Satzgrenzen. Auch andere über Textsorten hinweg anwendbare Kriterien, wie das Schriftbild oder die Textlänge, fallen in diese Kategorie. Textsortenspezifische Variablen beziehen sich auf die spezifischen Charakteristika einer der drei operationalisierten Textsorten (narrativ, argumentativ, informierend), wie im Falle der narrativen Texte beispielsweise die sprachliche Markierung des Wendepunktes. Da sich die Inhaltsvariablen von Aufgabe zu Aufgabe unterscheiden und auch innerhalb einer Textsorte nicht vergleichbar sind, wurden die Variablen für die Bewertung der Textinhalte aufgabenspezifisch entwickelt.

Je nach Thema und Charakter der Variablen wurden entweder dichotome Kodierungen (Codes 0; 1) oder niedrigstufige Ratings (Codes 1; 2; 3) gewählt. Alle analytischen Variablen für die Beispielaufgabe „Die zwei Esel“ sind in Tabelle II.4 zusammengetragen. Zur Veranschaulichung der konkreten Umsetzung dieses analytischen Kodierkonzeptes sind in Tabelle II.5 die Beschreibungen sowie Lösungs- und Nichtlösungsbeispiele für drei Inhaltsvariablen dokumentiert.

Tabelle II.4: Analytische Variablen der Aufgabe „Die zwei Esel“

Variablenlabel	Codes
Inhalt	
Darstellung von Inhaltsbereichen	
zentrale Figuren	0; 1
Verkettung der Esel	0; 1
Position der Futterhaufen	0; 1
Motivation zum Fressen/zur Annäherung	0; 1
Scheitern	0; 1
Ratlosigkeit	0; 1
Nachdenken	0; 1
Motivierung von Inhaltsbereichen	
Lösung	0; 1
Stimmigkeit	
Kohärenz	1; 2; 3
Narrative Elemente	
Struktur	
Überschrift	0; 1
Einführung neuer Aspekte (Exposition)	0; 1
Erzählabschluss	0; 1
Absätze	0; 1
Sprache – narrativ	
Markierung der Komplikation	0; 1
Markierung des Wendepunktes	0; 1
Markierung der Lösung	0; 1
Verknüpfung der Episoden	1; 2; 3
Gedanken und Emotionen der Akteure	1; 2; 3
Kommentare und Bewertungen der Schreibenden	0; 1
Erzählperspektive	0; 1
Sprache – allgemein	
Wortschatz	1; 2; 3
Wortwahl	1; 2; 3
Orthografie	1; 2; 3
Vielfalt der Satzkonstruktionen	1; 2; 3
Satzbau	1; 2; 3
Markierung von Satzgrenzen	1; 2; 3
Grammatikfehler	1; 2; 3
Wiederaufnahmen	1; 2; 3
Tempusgebrauch	1; 2; 3

Tabelle II.5: Original-Kodieranweisungen für drei Inhaltsvariablen der analytischen Kodierung für die Aufgabe „Die zwei Esel“

Inhalt		
Variablenlabel	**Beschreibung der Variablen und Vergabebedingungen der Codes**	**Lösungs- und Nichtlösungsbeispiele**
Zentrale Figuren	**Code 1:** Genannt werden mindestens Anzahl und Tierart der Akteure (zwei Esel) **Code 0:** Die unter Code 1 genannte Bedingung wird nicht erfüllt.	Code 1: „Zwei Esel“ „Zwei Esel stehen vor zwei Büschen.“ „Es warn mal zwei Esel.“ Code 0: keine Beispiele im Textkorpus „Die zwei Esel“
Verkettung der Esel	**Code 1:** Es wird dargestellt, dass die Esel aneinander gekettet / -gebunden sind: mit Strick, Schnur, Seil, Leine, Kette u. ä. *Erläuterung*: Auch eher unpassende Begriffe (wie z. B. verbunden, verknotet) gelten hier, solange eindeutig erkennbar ist, dass mit ihnen versucht wird, die Aneinanderkettung der Esel zu beschreiben. **Code 0:** Die unter Code 1 genannte Bedingung wird nicht erfüllt.	Code 1: „… sie sind durch ein seil verbunden …“ „…die aneinander gekettet wahren.“ „…weil sie miteinander festgebunde.“ „…und waren zusammen gebunden.“ „…und zwischen ihnen war eine Leine “ Code 0: „…weil sie angebunden sind.“ „Zwei Esel hatten was zuessen gesehen. Der eine Esel hatte was zuessen an der rechten seite gesehen und der andere Esel an der linken seite gesehen. Dann liefen sue loss der eine nach rechts und der andere nach liengs dann kriegten sie keine Luft mehr und der eine Esel hatte dann das essen fon den anderen gesehen und dan assen sie zusammen.“
Motivation zum Fressen / zur Annäherung	**Code 1:** Es wird erkennbar, dass die Esel zu den (Futter-)Haufen gelangen wollen. **Code 0:** Die unter Code 1 genannte Bedingung wird nicht erfüllt.	Code 1: „Als die beiden hunger bekahmen, wollte der eine Esel den einen haufen Futter und der andere den anderen haufen.“ „Esel Fritz und Esel Karl streiten sich. Sie haben beide großen Hunger. beide sind schon ganz nah am Busch.“ Code 0: „Zwei Esel wollen in zwei verschiedene Richtungen“

II.8.2 Holistische Kodierung

Die nachfolgend vorgestellte holistische Kodierung basiert auf einer von den Autoren vorgenommenen Übersetzung und Anpassung des NAEP Holistic Scoring Guide für narrative Aufgaben in der Version für die vierte Jahrgangsstufe (vgl. U.S. Department of Education, 2003).

Die Rater der Zusatzstudie wurden ergänzend zur Kodierung der analytischen Variablen instruiert, den *Gesamteindruck* jedes Textes auf einer Skala mit folgenden Ausprägungen zu bewerten:

(1) „ungenügend“,
(2) „mangelhaft“,
(3) „ausreichend“,
(4) „befriedigend“,
(5) „gut“,
(6) „sehr gut“

Die Beurteiler wurden darauf hingewiesen, dass die verbalen Bezeichnungen der Skalenausprägungen sowie die Tatsache, dass die Skala genau sechs Stufen aufweist, nicht im Sinne der deutschen Notengebung missverstanden werden dürfen, sondern gewählt wurden, um die Vergleichbarkeit mit der US-amerikanischen Originalskala gewährleisten zu können. Weiterhin wurden die Rater instruiert, bei ihrer Beurteilung der Schülertexte die folgenden Teilaspekte zu berücksichtigen und diese möglichst gleich gewichtet in ihr Urteil einfließen zu lassen:

- Beachten des Schreibanlasses (Erfüllen der Aufgabenstellung),
- Aufbau und Erzählstruktur,
- Sprachgewandtheit,
- Detailreichtum,
- Orthografie und Grammatik.

Bei Unsicherheiten in der Bewertung konnte ein Vergleich der vorliegenden Schülerantwort mit den jeweiligen Benchmarktexten angestellt werden. Wie bereits oben erwähnt, sind Benchmarktexte Schülerantworten, die als idealtypische Illustrationen der Leistung, die auf einer bestimmten Stufe einer Skala zu erwarten ist, verstanden werden können. Mitunter werden mit Benchmarktexten auch die Übergänge zwischen zwei Stufen illustriert, damit die Rater einen

Anhaltspunkt dafür gewinnen, welcher Text als Grenzfall gewertet werden muss und gerade noch der einen oder bereits der anderen Kategorie zugeordnet werden kann.

Neben dem sechsstufigen Gesamteindruck wurden drei weitere holistische Variablen geratet. Hierbei handelt es sich um vierstufige Variablen zu den Bereichen Stil, Inhalt und sprachliche Richtigkeit. Die Entscheidung für eine vierstufige Skala wurde nach einer Vorstudie mit fünf Ratern getroffen, die sowohl mit vier- als auch mit fünfstufigen Skalen gearbeitet haben. Eine sechsstufige Skala wie für den holistischen Gesamteindruck wurde von vornherein ausgeschlossen, da diese für die inhaltlich zwangsläufig enger gefassten Teilbereiche vermutlich eine zu starke Differenzierung dargestellt hätte.

Die fünfstufigen Skalen erzeugten bei den Ratern jedoch deutlich mehr Unsicherheit in der Bewertung und führten zu erheblich schlechteren Übereinstimmungen. Daher haben wir uns für den Einsatz von vierstufigen Skalen entschieden, obwohl die auf diese Weise erfasste Varianz eher geringer ist. Die Bereiche Inhalt, Stil und sprachliche Richtigkeit wurden aufgrund theoretischer Vorüberlegungen und nach eingehendem Literaturstudium gewählt (vgl. Glasswell, Parr & Aikmann, 2001). Von zentralem Interesse bei der Einführung dieser zusätzlichen holistischen Variablen für Facetten der Schreibkompetenz war die Frage, welche Teilaspekte der Schreibkompetenz sich am deutlichsten in einem holistischen Gesamteindruck niederschlagen. Auch die drei vierstufigen holistischen Variablen wurden, wie der sechsstufige Gesamteindruck, durch Kriterien definiert und mithilfe von Benchmarktexten veranschaulicht.

Tabelle II.6: Gesamteindruck – Narrative Aufgaben
(übersetzte und adaptierte Version des NAEP Holistic Scoring Guide; U.S. Department of Education, 2003)

Scores und Beschreibung	
6	Der Schülertext deutet auf eine sehr gute Schreibkompetenz in Bezug auf den Schreibanlass, kann aber wenige geringfügige Fehler beinhalten. Ein Text dieser Kategorie besitzt im allgemeinen folgende Eigenschaften: • zeigt einen sehr guten Aufbau mit klarer Erzählstruktur • enthält bemerkenswerte Details, die den erzählenden Charakter bereichern • demonstriert deutlich Sprachgewandtheit im Sprachgebrauch/Fähigkeiten im Umgang mit der Sprache • weist nahezu keine Fehler in Orthografie, Grammatik, Sprachgebrauch und Satzaufbau auf
5	Der Schülertext deutet auf eine gute Schreibkompetenz in Bezug auf den Schreibanlass, kann aber geringfügige Fehler beinhalten. Ein Text dieser Kategorie besitzt im allgemeinen folgende Eigenschaften: • zeigt im Aufbau eine klare Erzählstruktur • enthält Details, die den erzählenden Charakter bereichern • demonstriert erkennbare Sprachgewandtheit • weist wenige Fehler in Orthografie, Grammatik, Sprachgebrauch und Satzbau auf
4	Der Schülertext deutet auf eine befriedigende Schreibkompetenz in Bezug auf den Schreibanlass. Ein Text dieser Kategorie besitzt im allgemeinen folgende Eigenschaften: • ist adäquat aufgebaut, zeigt aber gelegentlich Schwächen in der Erzählstruktur • enthält Details, die dem Erzählten dienen • demonstriert angemessene Sprachgewandtheit • darf einige Fehler in Orthografie, Grammatik, Sprachgebrauch und Satzbau aufweisen, jedoch nicht als Häufung oder erkennbares Fehlermuster
3	Der Schülertext deutet auf eine ausreichende Schreibkompetenz in Bezug auf den Schreibanlass, ist aber offensichtlich fehlerhaft. Ein Text dieser Kategorie besitzt im allgemeinen folgende Eigenschaften: • ist ansatzweise aufgebaut, besitzt aber keine klare Erzählstruktur • enthält einige Details, die dem Erzählten dienen • demonstriert stark eingeschränkte Sprachgewandtheit • weist ein Muster oder eine Häufung von Fehlern in Orthografie, Grammatik, Sprachgebrauch oder Satzbau auf
2	Der Schülertext deutet auf eine eingeschränkte Schreibkompetenz in Bezug auf den Schreibanlass, ist stark fehlerbehaftet. Ein Text dieser Kategorie besitzt im allgemeinen folgende Eigenschaften: • zeigt keinen Aufbau und/oder keine Erzählstruktur • enthält wenige oder keine relevanten Details • zeigt schwere oder anhaltende Fehler in der Wortwahl und im sprachlichen Ausdruck • zeigt schwere Fehler in Orthografie, Grammatik, Sprachgebrauch oder Satzbau
1	Der Text deutet auf fundamentale Defizite der Schreibkompetenz hin. Ein Text dieser Kategorie besitzt im allgemeinen folgende Eigenschaften: • zeigt keinerlei Aufbau • ist nicht kohärent • enthält schwere und kontinuierlich auftretende Defizite in Orthografie, Grammatik, Sprachgebrauch oder Satzbau

II.9 Ergebnisse

Die Ergebnisdarstellungen gliedern sich in zwei Abschnitte. Zunächst wird über Beurteilerübereinstimmungen und Interraterreliabilitäten berichtet, anschließend über die Beziehung zwischen analytischen und holistischen Variablen. Letztere gliedert sich wiederum in Passagen zum inhaltlichen und zum allgemein sprachlichen Bereich.

II.9.1 Interraterreliabilität und Beurteilerübereinstimmung

Die Resultate werden im Folgenden für Gruppen von Variablen gleichen Datenniveaus vorgestellt. Zunächst werden Befunde für die dichotomen analytischen Variablen präsentiert und diskutiert, anschließend Resultate für die dreistufigen analytischen Variablen und abschließend für die mehrstufigen holistischen Urteile.

Die prozentualen Übereinstimmungen für die dichotomen Variablen der analytischen Kodierung des DPC, bei denen es sich um Variablen des Inhalts, der Struktur sowie des narrativen Sprachgebrauchs handelt, weisen eine Spannweite von .74 bis .98 bei einem Mittelwert von .86 auf, was auf den ersten Blick als erfreulich zu bewerten ist. Allerdings muss bei allen Variablen zusätzlich die Besetzung der Kategorien in Rechnung gestellt werden. Hier zeigt sich für einige dichotome analytische Variablen das Problem stark ungleicher Kategorienbesetzungen. Dies betrifft beispielsweise die Variable „Einführung der zentralen Figuren“, da diese Leistung von nur ca. drei Prozent der Schülerinnen und Schüler nicht erbracht wird. Somit ist die sehr hohe prozentuale Übereinstimmung von 98 Prozent, die für diese Variable erzielt wird, in erster Linie auf die ungleichen Kategorienbesetzungen zurückführbar (vgl. Abschnitt II.4).

Um derartige Interpretationsfehler zu vermeiden, scheint eine simultane Betrachtung von Cohens κ angezeigt. Für dieses Maß findet sich für die Gruppe der dichotomen analytischen Variablen ein Range von $\kappa = -.01$ bis $\kappa = .95$ bei einem mittleren Wert von $\kappa = .55$. Es ergeben sich also verschiedentlich Werte im inakzeptablen Bereich. Besonders unbefriedigende Ergebnisse resultieren für die Variablen „Erzählperspektive“ mit einem $\kappa = -.01$ und für die Variable „Kommentare und Bewertungen der Schreibenden“ mit einem $\kappa = .29$. Das inakzeptable Ergebnis der Kodierung der Erzählperspektive wird wiederum durch extrem ungleiche Kategorienbesetzungen verursacht. In diesem Fall weisen 99 Prozent aller Schülerinnen und Schüler die geforderte Kompetenz auf. Auch der erfasste Anteil wahrer Varianz der Schreibkompetenz, der bei null Prozent für die Erzählperspektive und bei zwei Prozent für die Kommentare und Bewertungen der Schreibenden liegt, sowie die ICC als mittlere Korrelation

der Rater von .00 beziehungsweise .03 legen nahe, auf beide Variablen zu verzichten.[5] Dies haben wir in nachfolgenden Studien getan.

Die dreistufigen analytischen Variablen, die in erster Linie dem allgemeinen sprachlichen Bereich entstammen, zeigen allein aufgrund ihrer mehrstufigen Natur geringere Übereinstimmungen als die dichotomen Variablen. Die prozentuale Übereinstimmung bewegt sich in einem Bereich von .39 bis .94 mit einem Mittelwert von .62. Damit erweist sich die prozentuale Übereinstimmung als oftmals zufriedenstellend und in vielen Fällen sogar als sehr gut, vereinzelt aber auch als inakzeptabel gering. Für Cohens κ ergibt sich eine Spannweite von $\kappa = -.02$ bis $\kappa = .52$ mit einem mittleren Wert von $\kappa = .23$.

Dieser Befund kann insgesamt nicht zufriedenstellen. Die Resultate für einzelne Variablen wie beispielsweise „Wiederaufnahmen“ mit einem $\kappa = -.02$ müssen als besonders inakzeptabel eingeschätzt werden. Auch die Befunde für die Intraklassenkorrelation, die sich in einem Range von acht bis 67 Prozent Varianzaufklärung mit einem Mittelwert von 37 Prozent bewegt, sind nicht durchgehend positiv. Es zeigt sich weiterhin, dass mithilfe der ICC die Tendenzen der anderen Maße der Übereinstimmung bestätigt werden. So weist auch hier wiederum die Variable „Wiederaufnahmen“ mit einer mittleren Raterkorrelation von .08 einen denkbar schlechten Wert auf.

Für die vierstufigen holistischen Variablen Inhalt, Stil und sprachliche Richtigkeit sowie den sechsstufigen Gesamteindruck ergeben sich mit .43 (Gesamteindruck) bis .49 recht geringe absolute Übereinstimmungen. Erfreulich sind jedoch die Resultate bei Hinzuziehung jeweils einer Nachbarkategorie. Dieses Übereinstimmungsmaß betrachtet also alle Fälle, bei denen die Rater um nicht mehr als eine Kategorie voneinander abweichen.

Die Resultate bewegen sich im Bereich von .90 (Gesamteindruck) bis .97. Für alle holistischen Variablen finden sich Intraklassenkorrelationen mit 68 bis 75 Prozent Varianzaufklärung. Diese Befunde können im Hinblick auf die in der einschlägigen Literatur berichteten Maße der Korrelation zwischen zwei Ratern als sehr zufriedenstellend angesehen und für den ersten Einsatz dieses holistischen Maßes im deutschsprachigen Grundschulbereich als Erfolg versprechender Zugang gewertet werden.

---

[5] Üblicherweise wird die Intraklassenkorrelation (ICC) nur für intervallskalierte Daten berechnet. Allerdings verwendet auch Brennan (2001) im Ansatz der Generalisierbarkeitstheorie Varianzkomponentenzerlegungen für dichotome und (wenige Abstufungen umfassende) ordinale Variablen. Die Varianzanteile werden hierbei deskriptiv interpretiert. Unter Rückgriff auf diese Überlegungen könnte auch im Rahmen von IRT-Modellen für kategoriale Variablen eine entsprechende Varianzzerlegung auf der Skala der latenten Fähigkeiten vorgenommen werden (vgl. hierzu Ansätze von Verhelst, 2001).

II.9.2 Vergleich von holistischer und analytischer Kodierstrategie

Befunde zum inhaltlichen Bereich

Für eine Klärung der in Abschnitt II.6 aufgeworfenen Frage, inwieweit holistische und analytische Kodierungen geeignet sind, dasselbe Konstrukt zu erfassen, werden zunächst Befunde für die Facette Inhalt vorgestellt. Hierfür werden in einem ersten Schritt die Inhaltsvariablen für beide Kodierstrategien (holistisch vs. analytisch) und beide Studien (am DPC in Hamburg sowie am IQB) extrahiert. Somit liegen uns Daten für zwei distinkte Ratergruppen vor, von denen die DPC-Rater ausschließlich die analytische Kodierstrategie einsetzten, die IQB-Rater aber sowohl analytisch als auch holistisch kodierten. Die Rater beider Studien kodierten dieselbe Stichprobe von 260 Schülertexten.

Eine Einschätzung des Inhalts gemäß der analytischen Kodierstrategie wurde aus der Summation der sieben dichotomen Einzelvariablen zur Charakterisierung des Inhalts separat für zwei Rater gewonnen. Für die Überprüfung der Eindimensionalität und somit zur Rechtfertigung der Summation der dichotomen Inhaltsvariablen der analytischen Kodierstrategie (DPC-Kodierung) wurde eine Hauptkomponentenanalyse der entsprechenden Variablen basierend auf tetrachorischen Korrelationen durchgeführt. Hierbei zeigte sich, dass die erste Hauptkomponente 60 Prozent der Varianz aufklärt und in einem Screeplot nur ein deutlicher Faktor erkennbar ist. Die Skalenbildung durch Summation der dichotomen Inhaltsvariablen der analytischen Kodierung erscheint somit gerechtfertigt. Für die holistische Einschätzung des Inhalts wurde auf die entsprechende vierstufige Variable zurückgegriffen.

Für die Beantwortung der Frage, wie hoch der Zusammenhang zwischen der analytischen und der holistischen Einschätzung des Inhalts ist, wurde ein Strukturgleichungsmodell berechnet, in dem um die Messfehler der Ratings korrigiert wurde. Hierfür wurden die analytischen Kodierungen der DPC-Rater mit den holistischen Kodierungen der IQB-Rater in Beziehung gebracht, um sicherzustellen, dass gefundene Beziehungen nicht auf Konsistenztendenzen identischer Rater in beiden Strategien zurückzuführen sind. Damit waren allerdings die Bedingungen für beide Kodierstrategien nicht absolut identisch, sondern es gab zwischen beiden Studien geringfügige Abweichungen in Schulung und Durchführung.

Wie im Abschnitt „Methodisches Vorgehen“ bereits näher ausgeführt, wurden die Rater paarweise auf Raterkombinationen und diese wiederum zufällig auf Schülertexte aufgeteilt. In den berichteten Strukturgleichungsmodellen treten also nicht zwei „reale“ Rater, sondern so genannte Pseudorater auf. Daher wird im Weiteren von einer Austauschbarkeit der Rater ausgegangen, weshalb Faktorladungen und Varianzen für beide Rater als identisch angenommen werden.

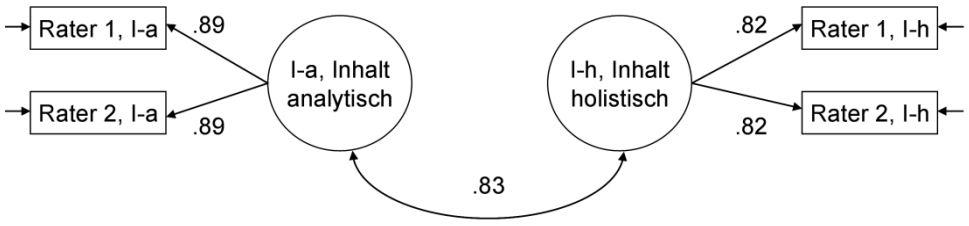

Abbildung II.3: Modell I: Zusammenhang der analytischen und holistischen Kodierung des Inhalts für zwei distinkte Ratergruppen aus separaten Studien

Anmerkung: Berichtet werden standardisierte Koeffizienten.
RMSEA = 0, SRMR = 0.03, CFI = 1; $\chi^2 = 2.65$, $df = 5$, $p = 0.753$

Für Modell I in Abbildung II.3 liegt die messfehlerbereinigte Korrelation zwischen den latenten Variablen für die mittels analytischer und holistischer Kodierung gewonnenen Einschätzungen des Inhalts bei $r = .83$. Dieser Befund könnte nun einerseits so gedeutet werden, dass die Inhaltsbewertungen in einer analytischen und einer holistischen Kodierung nicht als absolut identische Operationalisierungen ein und derselben Schülerfähigkeit zu verstehen sind. Eine andere mögliche Interpretation dieser Befundlage ist die, dass sich lediglich die Methodeneinflüsse verschiedener Zugänge bei der Abbildung ein und desselben Konstrukts niederschlagen. Dass die Korrelation zwischen beiden Inhaltsbewertungen von 1 verschieden ist, wäre demnach auf die spezifischen Kodiermethoden zurückzuführen und kein Beleg dafür, dass von separierbaren Konstrukten auszugehen ist. Letztere Interpretation scheint uns erkenntnislogisch plausibel. Bei Betrachtung der Faktorladungen fällt auf, dass die Kodierung gemäß der analytischen Strategie etwas höhere Reliabilitäten aufweist ($.89^2 = .79$) als die holistische Kodierung ($.82^2 = .67$).

Da davon auszugehen ist, dass bei absolut identischen Schulungen der Rater, genau gleichen Bedingungen und Umständen der Kodierung sowie gleichen Ratergruppen für beide Kodierstrategien tendenziell höhere Korrelationen resultieren sollten, wurde das eben vorgestellte Modell auf die Rater des IQB übertragen. Die in Abbildung II.4 dargestellten Zusammenhänge beziehen sich somit sowohl für die analytische[6] als auch für die holistische Beurteilung des

[6] In der Kodierung durch die IQB-Rater wurden nicht sieben, sondern nur sechs dichotome Variablen für die analytische Einschätzung des Inhalts kodiert und anschließend summiert, da die Variablen „D0962165 Ratlosigkeit“ und „D0962166 Nachdenken“ zusammengefasst worden waren.

Inhalts auf die Gruppe der IQB-Rater. Wiederum handelt es sich jeweils nicht um zwei reale, sondern um zwei austauschbare Pseudorater.

Wie der Abbildung II.4 entnommen werden kann, liegt die messfehlerbereinigte Korrelation zwischen der holistisch und der analytisch gewonnen Einschätzung des Inhalts für Modell II bei $r = .87$ und somit vernachlässigbar höher als für Modell I. Dies ist insofern bemerkenswert, als zu vermuten gewesen wäre, dass deutliche Halo-Effekte und somit höhere Korrelationen (vgl. Knoch, Read & von Randow 2007) auftreten, da ja nun beide Kodierstrategien von identischen Ratern umgesetzt wurden und Rater allgemein zu einer konsistenten Bewertung tendieren. Der Halo-Effekt als Residualkorrelation zwischen den Pseudoratern beider Kodierstrategien liegt jedoch nahe null ($r = .01$). Selbst nach Kontrolle des (nicht existenten) Halo-Effekts führt Modell II also zu einer sehr ähnlichen und nur geringfügig höheren Schätzung des Zusammenhangs zwischen den Resultaten beider Kodierstrategien für den Bereich Inhalt. Tatsächlich unterscheiden sich die gefundenen Korrelationen von $r = .83$ in Modell I und $r = .87$ in Modell II nicht signifikant ($p > .05$) voneinander, was bedeutet, dass die gefundene Differenz nicht gegen den Zufall abgesichert werden kann. Wiederum ergeben sich für die analytische Kodierung (Summenscore) geringfügig höhere Reliabilitäten ($.87^2 = .76$) als für die holistische Kodierung ($.82^2 = .67$).

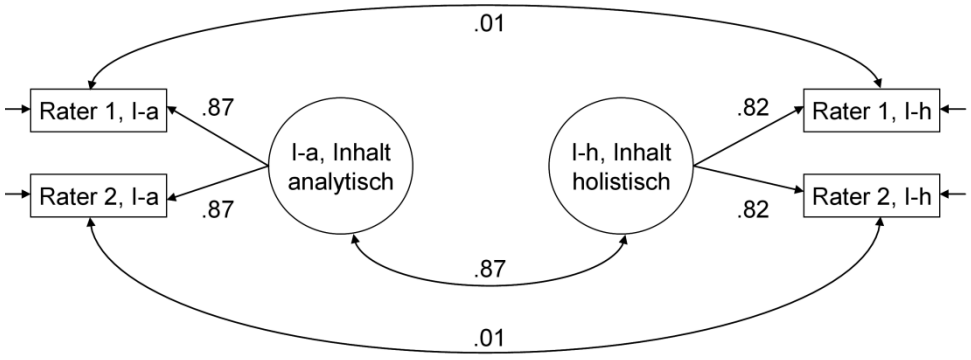

Abbildung II.4: Modell II: Zusammenhang der analytischen und holistischen Kodierung des Inhalts innerhalb der Rater einer Studie

Anmerkung: Berichtet werden standardisierte Koeffizienten.
RMSEA = 0, SRMR = 0.01, CFI = 1; $\chi^2 = 1.05$, $df = 4$, $p = 0.903$

Ermittelt man die Zusammenhänge des analytischen Summenscores für die Bewertung des Inhalts mit allen holistischen Variablen, so erhält man die in Tabelle II.7 dargestellten Ergebnisse.

Es zeigt sich, dass erwartungsgemäß die engste Beziehung zu einer holistischen Inhaltseinschätzung besteht. Die Höhe des Zusammenhangs repliziert das in Abbildung II.3 dargestellte Ergebnis. Überraschen mag der Befund, dass der analytisch gewonnene Summenscore in ähnlicher Höhe mit dem sechsstufigen Gesamteindruck, der die Inhaltsbeurteilung explizit berücksichtigt, korreliert wie mit der vierstufigen Einschätzung des Stils, die eine Berücksichtigung des Inhalts explizit ausschließt. Diese Befunde lassen vermuten, dass es den Ratern nicht gelingt, diese beiden Teilaspekte der Schreibkompetenz in ihrer Bewertung voneinander zu trennen. Wesentlich niedriger ist der Zusammenhang zwischen dem analytisch gewonnenen Inhaltseindruck und der vierstufigen holistischen Beurteilung der sprachlichen Richtigkeit. Hier gelingt den Beurteilern eine Trennung der Teilkompetenzen. Als Zwischenfazit zum Vergleich der holistischen und analytischen Kodierstrategie bei der Beurteilung des Inhalts eines narrativen Textes kann also festgehalten werden, dass sich ein hoher korrelativer Zusammenhang im Bereich von $r = .80$ bis .90 zeigt und hierin der Einfluss verschiedener Zugänge zur Abbildung eines identischen Konstrukts sichtbar wird.

Aufgrund der ermittelten Faktorladungen lässt sich ferner erkennen, dass die analytische Summenvariable geringfügig reliabler misst als die holistische Einschätzung des Inhalts. Dies entspricht bereits vorliegenden Befunden (vgl. Goulden, 1994).

Tabelle II.7: Messfehlerbereinigte Korrelationen zwischen dem analytischen Summenscore für die Beurteilung des Inhalts (DPC-Kodierung) und den holistischen Variablen (IQB-Kodierung) der Beispielaufgabe

	Gesamteindruck	Inhalt	Stil	sprachliche Richtigkeit
Summenscore Inhalt	.72	.83	.68	.31

Befunde zum allgemein sprachlichen Bereich

Auch für den allgemein sprachlichen Bereich soll ein Vergleich zwischen den Ergebnissen der analytischen und holistischen Kodierstrategie angestellt werden. Dieser Vergleich bezieht sich ähnlich wie das Modell I im vorhergehenden Abschnitt auf eine Gegenüberstellung der analytischen Kodierung der DPC-Rater und der holistischen Kodierung der IQB-Rater. In einem komplexen Strukturgleichungsmodell, welches der Logik einer Multi-Trait-Multi-Method-Modellierung (vgl. Eid, Nussbeck & Lischetzke, 2006) entspricht, wurden einerseits

Korrelationen zwischen den Faktoren beider Kodierstrategien (Tabelle II.8) und andererseits Halo-Effekte für beide Ratergruppen (Tabellen II.9 und II.10) bestimmt.

Die Tabelle II.8 zeigt zum einen die Zusammenhänge der allgemeinen Sprachvariablen der analytischen Kodierung (Kursivdruck) sowie die der vier holistischen Variablen (Fettdruck) untereinander. Zum anderen sieht man, inwieweit die allgemein sprachlichen Aspekte der analytischen Kodierung in den holistischen Variablen repräsentiert sind (grau unterlegter Bereich). Im linken oberen Bereich der Tabelle II.8 finden sich in Kursivdruck in erster Linie schwache bis mittlere positive Zusammenhänge. Lediglich die Variable „Tempusgebrauch" zeigt zu zwei anderen Variablen Nullzusammenhänge. Insgesamt weist dieses Korrelationsmuster auf eine deutliche Assoziiertheit der allgemein sprachlichen Aspekte hin, wobei alle Einzelvariablen unverkennbar voneinander trennbare Konstrukte darstellen.

Tabelle II.8: Messfehlerbereinigte Korrelationen zwischen den analytischen Variablen des allgemein sprachlichen Bereichs (DPC-Kodierung)* und den holistischen Variablen (IQB-Kodierung) der Beispielaufgabe

	Wort-schatz	Ortho-graphie	Vielfalt	Satz-bau	Satz-grenzen	Tempus	Gesamt-eindruck	Inhalt	Stil
Orthographie	*.50*								
Vielfalt Satz-konstruktion	*.53*	*.44*							
Satzbau	*.45*	*.58*	*.34*						
Markierung Satzgrenzen	*.19*	*.48*	*.24*	*.22*					
Tempus-gebrauch	*.15*	*-.02*	*.22*	*.15*	*.06*				
Gesamt-eindruck	.72	.58	.54	.42	.42	.08			
Inhalt	.52	.42	.42	.26	.30	.26	**.94**		
Stil	.80	.59	.64	.39	.38	.22	**.95**	**.88**	
Sprachliche Richtigkeit	.36	.82	.31	.45	.63	.15	**.67**	**.48**	**.65**

*Anmerkung: Um belastbare Aussagen generieren zu können, wurden in den hier berichteten Analysen nur solche allgemein sprachlichen Variablen berücksichtigt, die eine ICC von > .30 erreichten. Dies war für die Variablen Wortwahl, Grammatikfehler und Wiederaufnahme nicht der Fall.

Die Zusammenhänge der holistischen Variablen untereinander sind im rechten unteren Bereich der Tabelle II.8 in Fettdruck dargestellt. Diese liegen durchgehend im mittleren bis sehr hohen positiven Bereich, was eine sehr starke Assoziiertheit der vierstufigen holistischen Variablen für Inhalt und Stil mit dem sechsstufigen Gesamteindruck verdeutlicht. Die Beziehungen zu der vierstufigen Variablen für sprachliche Richtigkeit sind insgesamt geringer. Dies deutet auf eine Sonderstellung dieser Variablen innerhalb einer holistischen Aufsatzbewertung hin. Bemerkenswert scheint außerdem der hohe positive Zusammenhang zwischen den holistischen Einschätzungen für Inhalt und Stil in Höhe von $r = .88$, der aufgrund einer theoretisch plausiblen Differenzierbarkeit dieser beiden Teilaspekte der Schreibkompetenz und auch aufgrund der klar trennbaren Bewertungskriterien beider Variablen nicht unbedingt zu erwarten gewesen ist.

Die Befunde zur Dimensionalität des Konstrukts der Schreibkompetenz lassen auf der Grundlage holistischer Beurteilungen also einen engen Zusammenhang zwischen stilistischen und inhaltlichen Aspekten erkennen. Die Komponente der sprachlichen Richtigkeit (Rechtschreibung und Grammatik) lässt sich jedoch nicht ohne Weiteres als Bestandteil eines eindimensionalen Konstrukts fassen. Dies steht im Einklang mit Befunden zum Verfassen von Briefen in der Sekundarstufe I und II, wo durch Neumann (2007) eine ähnliche zweidimensionale Grundstruktur der Schreibfähigkeit identifiziert wurde.

Im grau unterlegten Bereich der Tabelle II.8 finden sich schließlich die Beziehungen zwischen analytischen und holistischen Variablen. Diese Befunde verdeutlichen, welches Gewicht die verschiedenen analytischen Sprachvariablen auf den jeweiligen holistischen Variablen haben. So erkennt man beispielsweise, dass die Variable „Orthografie“ erwartungsgemäß den mit Abstand höchsten Zusammenhang mit der sprachlichen Richtigkeit aufweist. Ferner wird deutlich, dass die allgemeinen Sprachvariablen der analytischen Kodierstrategie in ähnlichem Umfang im sechsstufigen holistischen Gesamteindruck und der vierstufigen holistischen Einschätzung des Stils repräsentiert sind, wobei einzelne analytische Sprachvariablen erwartungsgemäß höher auf dem Stil als auf dem Gesamteindruck laden. Überraschend erscheint wiederum das Beziehungsmuster mit der analytischen Variable „Tempusgebrauch“, welche den höchsten Zusammenhang nicht mit der Einschätzung des Stils oder dem Gesamteindruck, sondern mit der holistischen Bewertung des Inhalts aufweist. Dieser Befund verdeutlicht erneut den bereits thematisierten Problemgehalt dieser Variable.

In Ergänzung dieser Korrelationsanalysen wurde eine lineare Regression des holistischen Gesamteindrucks auf die analytischen Sprachvariablen auf manifester und latenter Ebene berechnet. Bei einer auf latenten Variablen basierenden Modellierung kann durch die Verwendung der analytischen Sprachvariablen als alleinige Prädiktoren eine Varianzaufklärung von $R^2 = .63$ erzielt werden. Sowohl in der Analyse auf manifester als auch in der auf latenter

Ebene erweist sich der Wortschatz als der mit Abstand wichtigste Prädiktor und besitzt im latenten Fall ein standardisiertes Regressionsgewicht von $\beta = .54$ ($p < .001$). Auch hier erweist sich also der Wortschatz als wesentlich, sowohl als Aspekt der allgemeinen Sprachvariablen in der analytischen Kodierung als auch als Bestandteil der holistischen Einschätzung des Stils.

Auch anhand der in Tabelle II.8 zusammengefassten Modellierung lässt sich feststellen, dass die allgemeinen Sprachvariablen der analytischen Kodierstrategie eher unreliabel gemessen werden. Dies äußert sich in der Höhe der standardisierten Messfehler, die für die analytischen Sprachvariablen im Bereich von .49 bis .73 und damit deutlich über den standardisierten Messfehlern der holistischen Variablen liegen, die sich zwischen .31 und .41 bewegen.

Neben den Zusammenhängen der Kodierstrategien wurden wiederum die Halo-Effekte für die allgemeinen Sprachvariablen der analytischen Kodierung sowie aller vier holistischen Variablen bestimmt. Die Ergebnisse dieser Analysen finden sich in den Tabellen II.9 und II.10. Wie Tabelle II.9 entnommen werden kann, ergeben sich für die analytischen Sprachvariablen durchgehend keine beziehungsweise lediglich sehr geringe Halo-Effekte. Die unabhängige Beurteilung der interessierenden Fähigkeiten gelingt also sehr gut.

Tabelle II.9: Korrelationen der Uniquenesses der DPC-Rater für die analytischen Variablen des allgemein sprachlichen Bereichs der Beispielaufgabe (Halo-Effekte)

	Wortschatz	Orthografie	Vielfalt Satz-konstruktion	Satzbau	Markierung Satzgrenzen
Orthografie	-.02				
Vielfalt Satzkonstruktion	.10	-.03			
Satzbau	.00	.06	-.02		
Markierung Satzgrenzen	-.07	.14	-.11	.09	
Tempusgebrauch	-.01	.21	-.07	.05	.18

Aus den in Tabelle II.10 für die holistischen Variablen zusammengefassten Befunden lässt sich unschwer erkennen, dass für die holistische Kodierstrategie ungleich größere Halo-Effekte identifiziert werden können. So zeigt der sechsstufige Gesamteindruck eine deutliche Überstrahlung der vierstufigen holistischen Einschätzungen für Inhalt, Stil und sprachliche Richtigkeit. Auch die Bereiche Stil und Inhalt weisen einen substanziellen Halo-Effekt auf.

Lediglich die Beurteilung der sprachlichen Richtigkeit scheint im Sinne der Einschätzung einer weitgehend differenzierbaren Subfähigkeit zu gelingen.

Tabelle II.10: Korrelationen der Uniquenesses der IQB-Rater für die holistischen Variablen der Beispielaufgabe (Halo-Effekte)

	Gesamteindruck	Inhalt	Stil
Inhalt	.36		
Stil	.47	.37	
Sprachliche Richtigkeit	.28	.23	.20

Insgesamt zeigen also die Variablen des allgemeinen Sprachbereichs der analytischen Kodierung deutlich geringere Halo-Effekte als holistische Einschätzungen der Schülertexte auf. Dies liegt zu einem gewissen Grad in der Natur der Variablen begründet – so fordert ja beispielsweise die Beurteilung des Gesamteindrucks die Integration aller als relevant identifizierter Teilkompetenzen – impliziert aber andererseits auch, dass eine differenzierte holistische Einschätzung von Facetten der Schreibkompetenz immer der Gefahr von unerwünschten Halo-Effekten ausgesetzt ist.

II.10 Diskussion und Ausblick

In diesem Beitrag wurde die Absicht verfolgt, unterschiedliche Strategien bei der Beurteilung von Schreibaufgaben vorzustellen und hinsichtlich ihrer Eignung in Large-Scale-Assessments zu analysieren. Dabei wurde zwischen einer analytischen und einer holistischen Bewertungsstrategie unterschieden und argumentiert, dass beide Strategien potenziell geeignet sind, zu validen Aussagen über Schreibkompetenzen zu gelangen. Wichtig erscheint uns in diesem Zusammenhang noch einmal der Hinweis, dass es erkenntnislogisch unplausibel ist, anzunehmen, mit beiden Strategien würden sich unterschiedliche Konstrukte erfassen lassen. Beide Strategien stellen vielmehr unterschiedliche Zugänge zur Messung desselben Personenmerkmals dar. Wir wollen im Folgenden zunächst methodische Aspekte diskutieren und anschließend Implikationen im Hinblick auf zukünftige Assessment-Studien vorstellen.

II.10.1 Methodische Aspekte

Bezug nehmend auf die Befunde zur Interraterreliabilität und Beurteilerübereinstimmung kann man pointiert festhalten, dass man mitunter zwar genau misst, wie im Falle der dichotomen analytischen Variablen, aber nicht immer zwischen guten und schlechten Schülern unterscheiden kann. Letzteres ist jedoch das eigentliche Ziel. Im Allgemeinen fallen Maße, wie die prozentuale Übereinstimmung oder Cohens κ, umso schlechter aus, je mehr Stufen die verwendete (Rating-)Skala umfasst. Entscheidend ist hierbei aber nicht die absolute, sondern die relative Übereinstimmung, da ausschlaggebend ist, dass verschiedene Rater die zu bewertenden Texte in eine annähernd gleiche Rangreihenfolge bringen.

Unserer Ansicht nach sollte daher das primäre Interesse dem Signal-Rausch-Verhältnis in den eingesetzten Variablen, also der Frage nach dem Verhältnis der Anteile von wahrer Varianz und Messfehlern, gelten und sich nicht allein auf eine weitgehend unreflektierte Bestimmung der prozentualen Übereinstimmung oder Cohens κ beschränken (vgl. Hayes & Hatch, 1999; vgl. Übersax, 2008). Diese Diskussion mündet letzten Endes in der Frage, ob angesichts absoluter Ungenauigkeit eine valide Messung möglich scheint oder ob hierfür Reliabilität zwingende Voraussetzung ist (vgl. LeBreton & Senter, 2008). Allerdings ist diese Frage stark vom jeweiligen Anwendungskontext bestimmt, also beispielsweise davon, ob Individualdiagnostik oder ein Bildungsmonitoring die Zielstellung ist.

Die Interpretation von Maßen zur Beurteilerübereinstimmung ist zudem abhängig davon, ob eine „wahre Kodierung" für ein Schülerprodukt als überhaupt möglich angenommen wird (vgl. Patz, Junker, Johnson & Mariano, 2002; vgl. Verhelst & Verstralen, 2001). Es muss in diesem Fall für jeden Schülertext für jede zu kodierende Variable genau eine „wahre Kategorie" geben, während Kodierungen anderer Kategorien demzufolge als „falsch" angenommen werden. Für dichotome Merkmale scheint in vielen Fällen die Existenz einer wahren Kategorie möglich, für (holistische) Ratingskalen ist dies aber kritisch zu hinterfragen. Dies entspricht beispielsweise auch der Modellierungstechnik der *Multifacettenmodelle* (vgl. Eckes, 2005), bei denen differenzielles Bewerterverhalten nicht auf den beobachteten Kategorien, sondern auf der Logitskala (auf der sich auch die Schülerfähigkeit befindet) modelliert wird. Ratereigenschaften beeinflussen demzufolge wie Items die Schätzung der Schülerfähigkeit auf der latenten Logitskala, die sich in einem probabilistischen IRT-Modell in manifestem Verhalten äußert. Bei Anwendung dieser Modellklassen wird demzufolge nicht von der Existenz einer wahren Kategorie ausgegangen beziehungsweise Abweichungen von dieser werden nicht manifest modelliert (siehe dazu im Gegensatz das *Hierarchical-Rater-Model*, Patz et al., 2002; für eine Gegenüberstellung dieser Modellansätze vgl. Böhme & Robitzsch, 2007). Generell soll erwähnt werden, dass bei

Multifacettenmodellen immer das *Repeated-Ratings-Problem* besteht (vgl. Patz et al., 2002; vgl. Verhelst & Verstralen, 2001; vgl. Wilson & Hoskens, 2001), bei dem Halo-Effekte im IRT-Modell ignoriert werden. Dies führt zu einer erhöhten Reliabilitätsschätzung der Schülerfähigkeiten (vgl. Patz et al., 2002). Insbesondere bei der Betrachtung der holistischen Variablen in diesem Beitrag wird jedoch deutlich, dass Halo-Effekte nicht vernachlässigbar sind.

Neben den genannten Varianzquellen zur Erfassung von mangelnder Reliabilität bei Ratings können Varianzkomponentenmodelle (vgl. Hoyt, 2000; vgl. Searle, Casella & McCulloch, 2006) – die in der Literatur auch in Modellen der Generalisierbarkeitstheorie auftreten – zur Erfassung möglicher Varianzquellen (z. B. Rater, Aufgabe, Aufgabentyp, Darbietungsmodus, Zeitpunkt des Ratings) eingesetzt werden. Diese Techniken erlauben oftmals eine präzisere Abschätzung von erwartbaren Messfehlern bei der Beurteilung von Schreibprodukten (vgl. Schoonen, 2005).

Es sollte festgehalten werden, dass analytische Variablen naturgemäß höhere Übereinstimmungen in Maßen wie der prozentualen Übereinstimmung oder Cohens κ erzielen, gleichzeitig aber in stärkerem Umfang dem Problem ungleicher Kategorienbesetzungen unterworfen sind und darüber hinaus meist weniger Varianz und damit auch weniger wahre Varianz in den Schülerantworten erfassen. Gleichzeitig zeigen analytische Variablen nur geringe Halo-Effekte und kleine Interkorrelationen. Dies bedeutet, dass eine Differenzierung von Teilaspekten der Schreibkompetenz im Fall der analytischen Kodierung sehr gut gelingt. Die holistische Kodierung zeigt in Varianzkomponentenzerlegungen sehr erfreuliche Ergebnisse im Hinblick auf die ermittelte wahre Varianz. Gleichzeitig schneidet die holistische Kodierung aufgrund der Vielstufigkeit der eingesetzten Skalen in Bezug auf Maße wie die prozentuale Übereinstimmung oder Cohens κ deutlich schlechter ab. Dies ist jedoch unserer Ansicht nach nicht primär relevant. Sinnvoller sind hier Maße der Interraterreliabilität wie die ICC, für die äußerst zufriedenstellende Ergebnisse gefunden wurden. Problematisch bleibt im Fall der holistischen Kodierstrategie allerdings das hohe Ausmaß an Interkorrelation der Teilbereiche, für die eine differenzielle Erfassung intendiert war. Ferner sind die starken Halo-Effekte ein Problem, dem man sich in künftiger Forschung stärker widmen sollte.

II.10.2 Implikationen für zukünftige Assessments

Die Frage, welcher Kodierstrategie in zukünftigen Studien aufgrund der vorliegenden Befunde der Vorzug zu geben ist, kann hier nicht abschließend beantwortet werden. Vielmehr wird man je nach Anlass und Zielstellung der Untersuchung beziehungsweise Testung abwägen müssen, mit welcher Strategie man den im Vordergrund stehenden Fragen beziehungsweise Intentionen

adäquater gerecht werden kann. Ein analytischer Kodierzugang besitzt aus fachdidaktischer Sicht zahlreiche wertvolle Vorteile, die sich insbesondere auf eine detaillierte Erschließung des Schreibproduktes beziehen, durch die sehr konkrete Aussagen über den Kompetenzstand eines Kindes getroffen werden können. In kleineren fachdidaktisch orientierten empirischen Untersuchungen zur Schreibkompetenz wird man daher dieser Strategie den Vorzug geben. Auch bei Fragen der Individualdiagnostik wird ein differenziertes, auf Teilkomponenten des Schreibproduktes fokussierendes Vorgehen bei der Bewertung die Methode der Wahl sein. Dem Problem, dass die für die allgemeinen analytischen Sprachvariablen derzeit vorliegenden Befunde zur Reliabilität nicht befriedigen können, sollte in der künftigen Forschung größere Aufmerksamkeit gewidmet werden.

Für die Hand von Lehrkräften, beispielsweise im Rahmen von Vergleichsarbeiten oder für die tagtägliche Korrekturarbeit, scheint die hier vorgestellte analytische Kodierung jedoch weniger geeignet zu sein. Zu umfangreich und auch fachlich zu komplex sind die einzelnen Kriterien gefasst. Zudem zeigen die Übereinstimmungsmaße, dass der hohe Zeitaufwand nicht etwa mit einer höheren Messgenauigkeit einhergeht, sondern vielmehr bei der Kodierung Unsicherheiten bleiben. Für den schulischen Kontext oder auch für die Korrektur von Schreibaufgaben im Rahmen von Vergleichsarbeiten bieten sich daher möglicherweise holistische Variablen für Teilbereiche der Schreibkompetenz wie die hier vorgestellten vierstufigen Variablen zu Inhalt, Stil und sprachlicher Richtigkeit an. Diese holistischen Bewertungen sollten aber keinesfalls als Ad-hoc-Eindrücke verstanden, sondern im Sinne wohldefinierter und kriterial verankerter Bewertungsmaßstäbe eingesetzt werden. Diese Art der Bewertung von Schreibprodukten könnte eine sinnvolle Balance zwischen notwendigen Informationen für eine fundierte Leistungsbeurteilung und gezielte Förderung einerseits und dem durch Lehrkräfte leistbaren Arbeitsaufwand andererseits darstellen. Zudem kann die Bereitstellung von Benchmarktexten und Beschreibungen der jeweiligen Stufen die Professionalität der Lehrkräfte im Bereich der Diagnostik von Schreibaufgaben befördern.

Im Rahmen großer Schulleistungsstudien schließlich, wie beispielsweise in den Ländervergleichen auf Basis der Bildungsstandards, die für den Primarbereich erstmalig im Jahr 2011 stattfinden werden, bietet sich vermutlich der Einsatz eines holistischen Gesamteindrucks neben einer geringen Anzahl analytischer und holistischer Variablen für Teilbereiche der Schreibkompetenz an. Ähnlich wie der hier vorgestellte sechsstufige Gesamteindruck, der sich an die Vorarbeiten aus NAEP (vgl. U.S. Department of Education, 2003) anlehnt, sollte ein solcher Gesamteindruck auf allen Stufen durch klare Kriterien charakterisiert sein und durch Schülerbeispiellösungen illustriert werden. Auf diese Weise ließen sich alle Aufgaben einer Textsorte – und möglicherweise sogar textsortenübergreifende Aufgaben – in einem

gemeinsamen Stufenmodell der Schreibfähigkeit abbilden. Nur durch einen einheitlichen, gemeinsamen Bewertungsmaßstab wäre außerdem die angestrebte Dokumentation von Entwicklungstrends leistbar.

Der mögliche Vorwurf, dabei bleibe unklar, welche Leistung auf Schülerseite eigentlich gemessen werde, kann anhand der Befunde aus Tabelle II.8 klar entkräftet werden. Man sieht dort sehr deutlich, woraus sich der Gesamteindruck speist. Es sind dies die vierstufigen holistischen Variablen Inhalt und Stil. Betrachtet man den Beitrag der allgemeinen analytischen Sprachvariablen, so rückt hier in erster Linie der Wortschatz als Aspekt des Stils in den Blick.

Nicht unerwähnt bleiben soll in diesem Zusammenhang das Erfordernis ökonomischer Erwägungen. Kodierungen in Large-Scale-Assessments finden notwendigerweise durch trainierte Rater statt, welche die Schülertexte mehrfach kodieren. Um hier bei Stichprobengrößen von ca. 40 000 Schülerinnen und Schülern den Zeitaufwand und die damit einhergehenden Kosten der Kodierung in einem vertretbaren Rahmen halten zu können, wird eine Beschränkung auf einen kriterial verankerten und durch Benchmarks illustrierten Globaleindruck sowie eventuell holistische Variablen für Teilbereiche der Schreibkompetenz möglicherweise der einzig gangbare Weg sein. Dass in Ergänzungsstudien oder für Teile der in Large-Scale-Assessments erhaltenen Schülertexte zusätzlich auch analytische Kodierungen unerlässlich sind, um offene Forschungsfragen bearbeiten zu können, wird hierdurch nicht in Frage gestellt.

Gewiss wird man bei der Beschränkung auf holistische Kodierungen im Rahmen großer Schulleistungsstudien mit der Frage konfrontiert werden, ob durch ein solches Vorgehen mit Blick auf den schulischen Alltag die richtigen Signale gesendet werden. Auch an dieser Stelle muss auf die jeweiligen Kontexte und Zielstellungen verwiesen werden, die hier deutlich voneinander abweichen. Was für den Einsatz in Large-Scale-Assessments adäquat erscheint, muss keinesfalls auch im Schulalltag der Weg der Wahl sein. Umso wichtiger ist es daher, in den flächendeckenden Vergleichsarbeiten und in der Fortbildung der Lehrkräfte ein Spektrum Erfolg versprechender diagnostischer Ansätze anzubieten.

II.11 Literatur

Augst, G. & Faigel, P. (1986). *Von der Reihung zur Gestaltung. Untersuchungen zur Ontogenese der schriftsprachlichen Fähigkeiten von 13 bis 23 Jahren*. Frankfurt am Main: Peter Lang.

Bachmann, T. (2002). *Kohäsion und Kohärenz. Indikatoren für die Schreibentwicklung*. Innsbruck: Studienverlag.

Bakeman, R. & Gottman, J. M. (1986). *Observing Interaction: An Introduction to Sequential Analysis*. Cambridge: Cambridge University Press.

Barret, S. (2001). The impact of training on rater variability. *International Education Journal*, *2*, 49-58.

Bates, D., Maechler, M. & Dai, B. (2008). *lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-26*.

Baurmann, J. (1995). Schreiben in der Schule. Orientierung an Schreibprozessen. In J. Baurmann, J. & R. Weingarten (Hrsg.), *Schreiben – Prozesse, Prozeduren und Produkte* (S. 51-69). Opladen: Westdeutscher Verlag.

Bechger, T. M., Maris, G. & Hsiao, Y. P. (2007). *Assessing the size of halo-effects in performance based tests and a practical solution to avoid halo-effects.* Measurement and Research Department Reports 2007-2. Arnheim: CITO, National Institute for Educational Measurement.

Becker-Mrotzek, M. (1997). *Schreibentwicklung und Textproduktion. Der Erwerb der Schreibtätigkeit am Beispiel der Bedienungsanleitung*. Opladen: Westdeutscher Verlag.

Becker-Mrotzek, M. & Böttcher, I. (2006). *Schreibkompetenz entwickeln und beurteilen*. Berlin: Cornelsen.

Bereiter, C. (1980). Development in Writing. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive Processes in Writing* (S. 73-93). Hillsdale, NJ: Erlbaum.

Bereiter, C. & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, NJ: Erlbaum.

Böhme, K. & Robitzsch, A. (2007). *Vergleich verschiedener IRT-basierter Skalierungsmodelle für die Textproduktion*. Vortrag auf der 8. Tagung der Fachgruppe Methoden und Evaluation der DGPs. 12.-14. September 2007, Gießen.

Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler* (6. Auflage). Berlin: Springer Verlag.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Bryant, B. R. & Bryant, D. P. (2003). Assessing the writing abilities and instructional needs of students. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of Psychological and Educational Assessment of Children: Intelligence, Aptitude and Achievement* (2nd Edition) (pp. 419-437). New York: Guilford Publications.

Congdon, P. J. & McQueen, J. (2000). The stability of rater severity in Large-Scale-Assessment Programs. *Journal of Educational Measurement*, *37*, 163-178.

Cooper, C. R. (1977). Holistic evaluation of writing. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 3-32). Urbana, IL: NCTE.

Eckes, T. (2005). Evaluation von Beurteilungen. Psychometrische Qualitätssicherung mit dem Multifacetten-Rasch-Modell. *Zeitschrift für Pychologie*, *213*, 77-96.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*, 155-185.

Eid, M., Nussbeck, F. W. & Lischetzke, T. (2006). Multitrait-Multimethod-Analyse. In F. Petermann & M. Eid (Hrsg.), *Handbuch der Psychologischen Diagnostik* (S. 332-345). Göttingen: Hogrefe.

Eigler, G., Jechle, T., Merziger, G. & Winter, A. (1990). *Wissen und Textproduzieren*. Tübingen: Narr.

Engelhard, G. (1994). Examining Rater Errors in the Assessment of Written Composition with a Many-Faceted Rasch Model. *Journal of Educational Measurement*, *31*, 93-112.

Feilke, H. (1996). Die Entwicklung der Schreibfähigkeiten. In H. Günther & O. Ludwig (Hrsg.), *Schrift und Schriftlichkeit. Writing and its Use. Ein interdisziplinäres Handbuch internationaler Forschung* (2. Halbband, S. 1178-1191). Berlin: deGruyter.

Feilke, H. (2003). Entwicklung schriftlich-konzeptualer Fähigkeiten. In U. Bredel, H. Günther, P. Klotz, J. Ossner & G. Siebert-Ott (Hrsg.), *Didaktik der deutschen Sprache* (Band 1, S. 178-192). Paderborn: Schöningh.

Feilke, H. & Schmidlin, R. (Hrsg.) (2005). *Literale Textentwicklung*. Frankfurt am Main: Peter Lang.

Fix, M. (2000). *Textrevisionen in der Schule*. Hohengehren: Schneider.

Fleiss, J. L. (1973). *Statistical Methods for Rates and Proportions*. New York: Wiley.

Fleiss, J. L. & Cohen, J. (1973). The equivalence of weighted Kappa and the Intraclass CorrelationCoefficient as measures of reliability. *Educational and Psychological Measurement, 33*, 613-619.

Fox, J. (2008). *sem: Structural Equation Models. R package version 0.9-12*.

Frick, T. & Semmel, M. I. (1978). Observer Agreement and Reliabilities of Classroom Observational Measures. *Review of Educational Research*, 48, 157-187.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2003). *Basian data analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Glasswell, K., Parr, J. & Aikmann, M. (2001). *Development of the asTTle Writing Assessment Rubrics for Scoring Extended Writing Tasks. Technical Report 6*. Auckland: University of Auckland.

Goldstein, H. (2003). Multilevel Statistical Methods. London: Arnold.

Goulden, N. R. (1994). Relationship of analytic and holistic methods to raters scores for speeches. *Journal of Research and Development in Education*, *27*, 73-82.

Grzesik, J. & Fischer, M. (1984). *Was leisten Kriterien bei der Aufsatzbeurteilung?* Opladen: Westdeutscher Verlag.

Harsch, C., Neumann, A., Lehmann, R. & Schröder, K. (2007). Schreibfähigkeit. In E. Klieme & B. Beck (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messungen. DESI-Studie* (S. 42-62). Weinheim: Beltz.

Hayes, J. R. & Hatch, J. A. (1999). Issues in measuring meliability. Correlation versus percentage of agreement. *Written Communication*, *16*, 354-367.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods*, *5*, 64-86.

Hug, M. (2001). *Aspekte zeitsprachlicher Entwicklung in Schülertexten. Eine Untersuchung im 3., 5. und 7. Schuljahr*. Frankfurt am Main: Peter Lang.

Huot, B. (1990). Reliability, Validity, and Holistic Scoring: What We Know and What We Need to Know. *College Composition and Communication*, *41*, 201-213.

Jechle, T. (1992). *Kommunikatives Schreiben. Prozess und Entwicklung aus der Sicht kognitiver Schreibforschung*. Tübingen: Narr.

Kahneman, D., Slovic, P. & Tversky, A. (1982). *Judgement under uncertainty. Heuristics and biases*. Hillsdale, NJ: Erlbaum.

KMK (2005): siehe Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2005).

Knoch, U., Read, J. & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, *12*, 26-43.

LeBreton, J. M. & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*, 815-852.

Lloyd-Jones, R. (1977). Primary trait scoring. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 33-66). Buffalo: State University of New York.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing*, *19*, 246-276.

Lumley, T. & McNamara, T. F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing*, *12*, 54-71.

Margraf, J. & Fehm, L. (1996). Methoden der klinischen Psychologie. In E. Erdfelder, R. Mausfeld, T. Meiser & G. Rudinger (Hrsg.), *Handbuch Quantitative Methoden* (S. 577-598). Weinheim: Beltz PVU.

Mullis, I. V. S., Martin, M. O. & Kennedy, A. M. (2004). *Item-writing guidelines for the PIRLS 2006 field test*. Presented at the 2nd PIRLS 2006 NRC Meeting Bratislava, Slovak Republic.

Myers, D. G. (2002): *Social psychology* (7th Edition). Holland, Michigan: Hope College.

Neumann, A. (2007). *Briefe schreiben in Klasse 9 und 11*. Münster: Waxmann.

Nitko, A. J. (2004). *Educational Assessment of Students*. Upper Saddle River, NJ: Pearson Prentice Hall.

Overall, J. E. & Magee, K. N. (1992). Estimating Individual Rater Reliabilities. *Applied Psychological Measurement*, *16*, 77-85.

Patz, R. J., Junker, B. W., Johnson, M. S. & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, *27*, 341-384.

R Development Core Team (2008). *R: A language and environment for statistical computing*. Wien: R Foundation for Statistical Computing.

Rupp, A. A., Vock, M., Harsch, C. & Köller, O. (2008). *Developing standards-based assessment tasks for english as a first foreign language. Context, processes, and outcomes in Germany*. Münster: Waxmann.

Schneuwly, B. (1988). *Le langage écrit chez l'enfant. La production des textes informatifs et argumentatifs*. Paris: Delachaux et Niestlé.

Schoonen, R. (2005). Generalizability of writing scores: an application of structural equation modeling. *Language Testing*, *22*, 1-30.

Searle, S. R., Casella, G. & McCulloch, C. E. (2006). *Variance Components*. New York: Wiley.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2005). *Bildungsstandards im Fach Deutsch für den Primarbereich (Jahrgangsstufe 4) – Beschluss vom 15.10.2004*. München: Wolters Kluwer.

Sieber, P. (2003). Modelle des Schreibprozesses. In U. Bredel, H. Günther, P. Klotz & G. Siebert-Ott (Hrsg.), *Didaktik der deutschen Sprache. Ein Handbuch* (Band 1, S. 208-223). Paderborn: Schöningh.

Slater, S. C. & Boulet, J. R. (2001). Predicting holistic ratings of written performance assessments from analytic scoring. *Advances in Health Sciences Education*, *6*, 103-119.

Steffen, K. (1995). *Schreibkompetenz. Schreiben als intelligentes Handeln*. Hildesheim: GeorgOlms.

Thorndike, E. L. (1920). A constant error on psychological rating. *Journal of Applied Psychology*, *IV*, 25-29.

Persky, H. R., Daane M. C. & Jin, Y. (2003). *The Nation's Report Card: Writing 2002*. Washington, DC: U.S. Department of Education. Institute of Education Sciences. National Center for Education Assessment (NCES).

Übersax, J. S. (2008). *Statistical Methods for Rater Agreement*. Zugriff am 31.05.2009 unter www.ourworld.compuserv.com/hompages/jsuebersax.

U.S. Department of Education. Institute of Education Sciences. National Center for Education Assessment (2003). *The Nation's Report Card: Writing 2002. NCES 2003–529*. Washington, DC.

Verhelst, N. D. & Verstralen, H. F. M. (2001). An IRT model for multiple raters. In A. Boomsma, M. A. J. Van Dujin & A. B. Snjiders (Eds.), *Essays on item response modeling* (pp. 89-108). New York: Springer.

Weigle, S.C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*, 263-287.

Weigle, S.C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.

Wilson, M. & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics, 26*, 283-306.

Wirtz, M. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität*. Göttingen: Hogrefe.

3.3.2 Richtig schreiben

Unter der Bezeichnung *richtig schreiben* ist neben der Textproduktion auch die orthografische Kompetenz der Schülerinnen und Schüler Bestandteil des Kompetenzbereichs Schreiben. Zusätzlich zu den hier aufgeführten Bestandteilen sind die Kompetenzaspekte der Zeichensetzung und – durch die Behandlung der Wortarten – der Groß- und Kleinschreibung im Kompetenzbereich *Sprache und Sprachgebrauch untersuchen* verortet.

Durch die Einbeziehung der orthografischen Kompetenz in den Bereich *Schreiben* wird eine integrative Sichtweise deutlich, welche darauf abzielt, das orthografisch und grammatikalisch korrekte Schreiben als Bestandteil des Verfassens und Überarbeitens von Texten zu verstehen. So sinnvoll und wünschenswert diese Sichtweise für die unterrichtliche Praxis ist, so problematisch ist sie für die testdiagnostische Umsetzung (vgl. Granzer, Böhme & Köller, 2008). Die Diagnose der orthografischen Kompetenz erfolgte daher in separaten Testinstrumenten. Im Folgenden soll diese Entscheidung kurz begründet werden.

Wie bereits oben angedeutet, unterscheiden sich die von den Schülerinnen und Schülern verfassten Texte beträchtlich hinsichtlich ihrer Länge und ihrer sprachlichen Komplexität. Dies betrifft auch die in Texten realisierten orthografischen Phänomene und verwendeten grammatikalischen sowie Satzkonstruktionen. Verschiedene Kinder, die in ihren Texten gleich viele orthografische und grammatikalische Fehler machen, verfügen daher nicht zwingend über dasselbe Maß an Kompetenz. In Abhängigkeit vom Umfang sowie der Bekanntheit und Komplexität der in den Texten gewählten Wörter müssten die *gemachten Fehler* hinsichtlich der Zahl und Qualität der *möglichen Fehler* relativiert werden. Dies ist jedoch nicht möglich. Zunächst wäre ein solches Vorgehen extrem zeitaufwändig, da es die genaue Analyse jedes Schülertextes hinsichtlich der Type-Token-Ratios von Wörtern und Fehlern unter zusätzlicher Berücksichtigung von Bekanntheit und Schwierigkeit der Wörter und grammatikalischen Konstruktionen erfordern würde. Tatsächlich ist letzteres aber ausgeschlossen, da jedes Bundesland und mitunter sogar jede Schule und Klasse eigene Wortlisten hat, die geübt und somit von den Schülerinnen und Schülern mit größerer Sicherheit „beherrscht“ werden. Solche häufig geübten Wörter für jedes Kind zu identifizieren, ist nicht möglich. Daher wurde zwar die sprachliche Richtigkeit bei der Beurteilung der Schreibprodukte berücksichtigt (vgl. Abschnitt II), primär wurden für die Ermittlung der orthografischen Kompetenz aber spezifische Testinstrumente eingesetzt, die speziell für dieses Kompetenzkonstrukt entwickelt wurden.

Neben der testmethodischen Begründung, die orthografische Kompetenz der Schülerinnen und Schüler vorrangig separat und nicht integriert bei der Untersuchung der

Textproduktion zu erfassen, findet diese Entscheidung auch dadurch empirische Unterstützung, dass sie in Strukturanalysen als eigenständiges Konstrukt ausgewiesen werden kann. Bremerich-Vos, Böhme und Robitzsch (2009) prüften die Struktur der sprachlichen Kompetenzen im Primarbereich und untersuchten hierbei auch die Zusammenhänge zwischen dem freien Schreiben und der orthografischen Kompetenz. In einer gemeinsamen Skalierung dieser beiden Kompetenzbereiche ergab sich eine messfehlerbereinigte Korrelation in Höhe von $r = .83$ (Bremerich-Vos, Böhme & Robitzsch, 2009, S. 210). Dieser Zusammenhang ist zwar enger als die Beziehungen aller anderen Kompetenzbereiche zum Schreiben, liegt aber in der gleichen Größenordnung wie die Korrelation zwischen den Kompetenzbereichen Lesen und Zuhören, die $r = .85$ beträgt (Bremerich-Vos, Böhme & Robitzsch, 2009, S. 210). Wertet man die Höhe der hier berichteten korrelativen Zusammenhänge als Hinweis auf die dimensionale Struktur[1], so könnte man argumentieren, dass bei einer Zusammenfassung der Kompetenzbereiche Schreiben und Rechtschreibung auch von der Identität des Lesens und des Zuhörens ausgegangen werden müsste. Da letzteres zwar diskutiert, aber sowohl aus psychologischer als auch aus deutschdidaktischer Sicht abgelehnt wird, sollten auch das Schreiben und die orthografische Kompetenz als separate Konstrukte verstanden werden. Für diese Sicht sprechen ferner differentielle Zusammenhänge der klassenzentrierten Deutschnoten mit den latenten, im Test ermittelten Kompetenzständen in den Bereichen Schreiben und Rechtschreibung. Diese betragen für Schreiben $r = –.65$, für die orthografische Kompetenz hingegen $r = –.73$ (Bremerich-Vos, Böhme & Robitzsch, 2009, S. 214). Somit zeigt die orthografische Kompetenz den engsten Zusammenhang mit der Benotung im Fach Deutsch. Dieser ist noch stärker als die Beziehung zwischen der Deutschnote und der Lesekompetenz und deutlich größer als der Zusammenhang der Deutschnote zum Schreiben. Dieser differentielle Befund spricht somit ebenfalls dafür, die orthografische oder Rechtschreibkompetenz der Grundschulkinder als eigenständigen Kompetenzbereich zu behandeln.

Die Untersuchung dieses Kompetenzbereichs ist Gegenstand des folgenden empirischen Beitrags:

[1] Zu einer kritischen Sicht auf eine solche Interpretation vgl. Böhme & Robitzsch (2009a).

III. Beitrag 3: Diagnostik der Rechtschreibkompetenz in der Grundschule – Konstruktprüfung mittels Fehler- und Dimensionsanalysen

Autoren:
Katrin Böhme / Albert Bremerich-Vos

Erschienen in:
D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.) (2009). *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 330-356). Weinheim: Beltz.

Für die Evaluierung der Bildungsstandards im Bereich der Rechtschreibkompetenz in der Grundschule wählten wir einen diagnostischen Ansatz, der sowohl quantitative Aussagen über das globale Kompetenzniveau als auch Aussagen auf der Ebene von qualitativen Fehleranalysen innerhalb der orthografischen Stufe des Rechtschreiberwerbs gestattet. Die Fehleranalysen basieren auf neun orthografietheoretisch plausiblen Fehlerkategorien, die in Anlehnung an die Aachener Förderdiagnostische Rechtschreibfehler-Analyse AFRA (vgl. Herné & Naumann, 2005) gewählt wurden.

Empirisch zeigt sich, dass eine Diagnostik der globalen Rechtschreibfähigkeit unter Berücksichtigung der Anzahl richtig geschriebener Wörter zum Ende der dritten und vierten Jahrgangsstufe geleistet werden kann. Eindimensionale Rasch-Skalierungen auf Wortebene liefern für alle vier entwickelten Testformen im Hinblick auf die Modellpassung gute Ergebnisse. Im Vergleich der Jahrgangsstufen 3 und 4 zeigt sich eine mittlere Leistungsdifferenz von $d = 0.70$, wobei zum Ende der vierten Jahrgangsstufe leichte Deckeneffekte auftreten. Erwartungsgemäß weisen Mädchen in beiden Jahrgangsstufen einen Leistungsvorsprung gegenüber den Jungen auf. Im Mittel ergibt sich hier ein Effekt von $d = 0.20$.

Auf deskriptiver Ebene fällt die überwiegende Mehrzahl der Fehler in die neun definierten Fehlerkategorien, womit unterstrichen wird, dass die Diagnostik innerhalb der orthografischen Stufe dem tatsächlichen Erwerbsstand entspricht. Eine Prüfung der Mehrdimensionalität auf Fehlerebene gemäß den neun orthografietheoretisch definierten Fehlerkategorien mittels einer Hauptkomponentenanalyse der tetrachorischen Korrelationen, mehrdimensionaler Rasch-Skalierungen sowie mithilfe kognitiver Diagnosemodelle ergibt keine Evidenz für neun Dimensionen. Theoriegeleitete und explorative Analysen legen eine eindimensionale Lösung nahe.

III.1 Überblick

Die hier vorgenommene Überprüfung der Rechtschreibkompetenz ist zum einen auf die länderübergreifenden Bildungsstandards bezogen (vgl. KMK, 2005a), zum anderen auf aktuelle fachdidaktische Überlegungen, die in den Standards nicht explizit aufgegriffen werden. Dies betrifft Entwicklungsmodelle der Rechtschreibkompetenz (vgl. Frith, 1986; vgl. Scheerer-Neumann, 1998, 2003), die Unterscheidung von Basis- und Orthographemen (vgl. Herné & Naumann, 2005; vgl. Thomé & Thomé, 2004), die Rolle der Silbe im Rechtschreiberwerb (vgl. Eisenberg, 2005; vgl. Ossner, 1996; vgl. Röber-Siekmeyer & Tophinke, 2002), die Unterscheidung von „Nachdenk"- und „Merkwörtern" (vgl. Scheerer-Neumann, 1986) sowie die Beschreibung von Fehlerlupenstellen für diagnostische Zwecke (vgl. Herné & Naumann, 2005;

vgl. May, 2002; vgl. Thomé & Thomé, 2004). Darüber hinaus sind methodische beziehungsweise testdiagnostische Erwägungen von Bedeutung, welche sich auf die optimale Operationalisierung der Rechtschreibkompetenz, ihre Ausdifferenzierung in Teilkompetenzen und die empirische Sicherung dieser Dimensionalitätsannahmen beziehen (vgl. Klieme et al., 2003). Im Mittelpunkt dieses Kapitels steht somit die Frage, inwieweit theoretisch abgeleitete didaktische Annahmen durch empirische Befunde gestützt werden können. In einer längerfristigen Perspektive soll so auch zu einer inhaltlich-theoretischen Anreicherung und verbesserten testdiagnostischen Überprüfung der Bildungsstandards beigetragen werden.

Im Folgenden wird zunächst in den Abschnitten III.2 und III.3 ein Überblick über das didaktisch fundierte Konstrukt der Rechtschreibkompetenz und zu Modellen des Rechtschreiberwerbs gegeben. In Abschnitt III.4 geht es um die in den Bildungsstandards verankerten Kompetenzerwartungen im Bereich der Rechtschreibung und darum, wie hier testdiagnostisch verfahren werden kann. In Abschnitt III.5 werden diagnostische Möglichkeiten und bereits vorliegende Testverfahren, insbesondere unter Berücksichtigung qualitativer Fehleranalysen, vorgestellt. Nach einer Präzisierung der Fragestellungen im Kontext des Ziels, die Bildungsstandards zu evaluieren, wird in Abschnitt III.7 die Anlage der Untersuchung vorgestellt, wobei die Testentwicklung, die diagnostischen Fehlerkategorien, die Datengrundlage sowie die statistischen Analysen beschrieben werden. Die in Abschnitt III.8 dargestellten Ergebnisse beziehen sich auf deskriptive Befunde, Geschlechts- sowie Jahrgangsdifferenzen und Resultate der für die Klärung der Konstruktdimensionalität durchgeführten Analysen.

III.2 Rechtschreibkompetenz

Die Orthografie ist ein System kodifizierter Normen der Verschriftung von (mündlicher) Sprache mithilfe von Buchstaben (genauer: Graphemen) und anderen Zeichen, insbesondere Satzzeichen. In ihr sind grammatische Informationen kodiert, zum Beispiel in Form der Großschreibung, mit der man (neben anderem) substantivische Kerne von Nominalphrasen und den Satzbeginn anzeigt. Alle alphabetischen Schriften und so auch das Deutsche basieren zu einem gewissen Grad auf der Korrespondenz von Lauten und Buchstaben beziehungsweise – präziser – von Phonemen und Graphemen. Das Gesprochene ist allerdings ein „Strom“ und keine Reihe per se distinkter Laute. Die Fähigkeit, einzelne Laute auszugliedern, entwickelt sich im Wesentlichen im Rahmen des schulischen Schriftspracherwerbs, und was zu verschriften ist, kann man aus „genauem“ Hören auf die Standard- oder dialektale Lautung allein oft nicht erschließen. So erscheint, was als *Vogel* zu schreiben ist, mündlich als [fo:gl], d. h. nicht als Zwei-, sondern als Einsilber. Wer hier korrekt schreibt, bezieht sich auf eine von der Schriftsprache her zu

begreifende Explizitlautung, deren Merkmale Eisenberg (2005) im Einzelnen beschrieben hat. Die Diagnostik der Rechtschreibkompetenz hat sich also stets sowohl auf lautanalytische als auch auf grammatische Aspekte zu beziehen (vgl. Valtin, Badel, Löffler, Meyer-Schepers & Voss, 2003), wobei sich die Gewichtung je nach Erwerbsphase verschiebt.

III.3 Erwerb der Rechtschreibkompetenz

Im Rahmen der Orthografiedidaktik arbeitet man seit geraumer Zeit mit Modellen von „Stufen" oder auch „Phasen" orthografischer Kompetenz (vgl. Brügelmann & Brinkmann, 1994; vgl. Frith, 1986; vgl. Günther, 1986; vgl. Scheerer-Neumann, 2003; vgl. Valtin, 1988). Besonders bekannt ist die folgende Stufung:

1. *präliteral-symbolisch:* Die Kinder haben verstanden, dass Zwei- für Dreidimensionales stehen kann, zum Beispiel Fotografien für Menschen (insofern handelt es sich eigentlich um eine für den Schriftspracherwerb unspezifische Vorläuferfähigkeit). Sie kritzeln und produzieren grafische Gestalten, mit denen sie Schrift imitieren und die sie als Nachahmungen von Schreibbewegungen Erwachsener ansehen (vgl. Günther, 1986).
2. *logographemisch*: Die Kinder orientieren sich an einzelnen, visuell hervorstechenden Merkmalen, zum Beispiel an der typografischen Form von Logos oder dem Anfangsbuchstaben des eigenen Namens. Andere Segmente werden gar nicht oder variabel, womöglich aber auch – auf der Basis einer „ganzheitlichen" Merkstrategie – richtig verschriftet. Ob es überhaupt sinnvoll ist, nicht nur für das Lesen, sondern auch für das Schreiben eine logographemische Stufe anzusetzen, ist strittig (vgl. Eichler, 1986). Mitunter wird die Phase, in der erste Schreibversuche mit lautlichem Bezug erkennbar sind, auch als protoalphabetisch-phonetische Phase bezeichnet (vgl. Thomé, 2003).
3. *alphabetisch*: Die Zerlegbarkeit des Lautstroms und die Links-rechts-Sequenzierung werden von den Kindern erkannt. Aus dem Kontinuum können auch Einheiten unter- und oberhalb der Silbe (Phoneme, Wörter) ausgegliedert und mit Graphemen verknüpft werden.
4. *orthografisch*: Dehnung, Schärfung, Umlautung, Auslautverhärtung und „Merkelemente" wie bei *Vater, Hexe* usw. werden von den Kindern zunehmend bedacht. Sie orientieren sich mehr und mehr am silbischen und morphematischen Prinzip, sodass sich zum Beispiel morphematische Übergeneralisierungen wie *vertig* für *fertig* finden lassen.

Die Annahme, dass es sich bei den hier beschriebenen Phasen um klar abgrenzbare und im Erwerbsprozess strikt aufeinanderfolgende Entwicklungsstufen handelt, wird mittlerweile sowohl

in der internationalen (Lennox & Siegel, 1998) als auch der nationalen Literatur (vgl. Eichler, 1986; vgl. May, 2002; vgl. Thomé, 2003) abgelehnt. Die Vorstellung einer strikten Stufenabfolge wird zunehmend abgelöst durch das Konzept parallel verfügbarer, sich nach und nach entfaltender Zugriffsweisen beziehungsweise Strategien.

In diesem Sinne grenzt May (2002) die Begrifflichkeiten „Stufe“ und „Strategie“ voneinander ab:

> Zwar werden die weiter reichenden Strategien nach den empirischen Befunden gewöhnlich später von den Lernenden aufgegriffen und realisiert, aber schon nach relativ kurzer Lernerfahrung greifen die Kinder normalerweise auf verschiedene Strategien zurück […]. Diese grundlegenden Rechtschreibstrategien sind deshalb auch keine ‚Stufen‘ in dem Sinne, dass ein Kind von einer Stufe zur nächsten voranschreitet. (May, 2002, S. 143).

III.4 Rechtschreibkompetenz in den Bildungsstandards und Möglichkeiten ihrer Testung

In den von der Kultusministerkonferenz verabschiedeten Bildungsstandards im Fach Deutsch für den Primarbereich heißt es im Hinblick auf die orthografischen Kompetenzen der Schülerinnen und Schüler am Ende der vierten Jahrgangsstufe:

> Die Kinder verfügen über grundlegende Rechtschreibstrategien. Sie können lautentsprechend verschriften und berücksichtigen orthografische und morphematische Regelungen und grammatisches Wissen. Sie haben erste Einsichten in die Prinzipien der Rechtschreibung gewonnen. Sie erproben und vergleichen Schreibweisen und denken über sie nach. Sie gelangen durch Vergleichen, Nachschlagen im Wörterbuch und Anwenden von Regeln zur richtigen Schreibweise. Sie entwickeln Rechtschreibgespür und Selbstverantwortung ihren Texten gegenüber. (KMK, 2005a, S. 8)

Unter der Überschrift „richtig schreiben“ werden die folgenden Kompetenzaspekte benannt[1]:

- geübte, rechtschreibwichtige Wörter normgerecht schreiben;
- Rechtschreibstrategien verwenden: Mitsprechen, Ableiten, Einprägen;
- Zeichensetzung beachten: Punkt, Fragezeichen, Ausrufezeichen, Zeichen bei wörtlicher Rede;
- über Fehlersensibilität und Rechtschreibgespür verfügen;
- Rechtschreibhilfen verwenden: Wörterbuch nutzen; Rechtschreibhilfen des Computers kritisch nutzen;
- Arbeitstechniken nutzen: methodisch sinnvoll abschreiben; Übungsformen selbstständig nutzen, Texte auf orthografische Richtigkeit überprüfen und korrigieren.

Bedenkt man, welche der hier genannten Teilkompetenzen einer Testung am ehesten zugänglich sind, dann ergibt sich folgendes Bild: Die Fähigkeit, die genannten Hilfsmittel zu nutzen, ist aus technischen Gründen nur eingeschränkt überprüfbar. Man könnte Ausschnitte aus Wörterbüchern verwenden beziehungsweise selbst konstruieren und auf dieser Basis untersuchen, inwiefern die Schülerinnen und Schüler einfache Suchstrategien beherrschen.[2] Die Fehlersensibilität und das „Rechtschreibgespür“ könnten über Korrekturaufgaben zugänglich werden, die ja auch unter dem Titel „Arbeitstechniken nutzen“ explizit genannt sind. Hier könnten auch die in den Standards genannten Aspekte der Zeichensetzung eine Rolle spielen; sie wären allerdings ebenfalls als Elemente von Schreibaufgaben denkbar.

Wir haben uns dafür entschieden, die ersten beiden Substandards ins Zentrum zu rücken, d. h. zu untersuchen, inwieweit „rechtschreibwichtige“ Wörter normgerecht geschrieben werden können, ob man sich Schreibungen erfolgreich „eingeprägt“ hat beziehungsweise ob man sie auf dem Weg des „Ableitens“ erschließen kann. Ob die Schreibung bestimmter Wörter im Unterricht thematisiert wurde, lässt sich nicht ermitteln, allenfalls anhand der Analyse von länderspezifischen Grundwortschätzen vermuten.

---

[1] Es soll hier nicht darum gehen, die Standards im Einzelnen zu kritisieren. Angesichts dessen, was derzeit im rechtschreibdidaktischen Diskurs erörtert wird, muten sie recht vage an, was aber intendiert war, da man die Präzisierung zukünftigen Arbeiten überlassen wollte. Besonders instruktiv ist der Vergleich mit den Standards für den Hauptschul- und den Mittleren Bildungsabschluss. Man kann den Eindruck gewinnen, dass die Schülerinnen und Schüler im Lauf der Sekundarstufe I nicht mehr sehr viel hinzulernen müssen beziehungsweise sollen (KMK, 2004, 2005b).

[2] Diese Operationalisierung wurde ebenfalls gewählt. Über Resultate berichten wir in Bremerich-Vos & Böhme (2009).

III.5 Diagnostische Möglichkeiten im Bereich der Rechtschreibung

Im Rahmen standardisierter Rechtschreibtests haben die Schülerinnen und Schüler in der Regel nach Diktat einzelne Wörter oder Sätze zu schreiben oder Lückensätze beziehungsweise -texte nach Diktat zu ergänzen. Hinzu kommt als Variante die Korrektur von Falschschreibungen (vgl. Herné, 2003). Hierbei geht es aber eher um Rechtschreibwissen als um Rechtschreibkönnen. In dem Maß, in dem im Unterricht verstärkt auf eine Orientierung am Prozess (des Schreibens) gesetzt wird, bei dem das Überarbeiten eigener und fremder Texte eine große Rolle spielt, gewinnt die Testform „Korrektur von Falschschreibungen" jedoch an ökologischer Validität.

Der Einsatz von Lückentexten beziehungsweise Lückensätzen hat aus unserer Sicht folgende Vorteile:

- Unterschiede im Schreibtempo der Schülerinnen und Schüler fallen kaum ins Gewicht;
- der Schreibaufwand und damit der zeitliche Umfang der Testung sind begrenzt;
- das Schreiben von sicher beherrschten Wörtern wie Artikeln und Pronomen entfällt und
- die Aufmerksamkeit kann ungeteilt der Rechtschreibung gelten und wird nicht, wie zum Beispiel bei Textdiktaten, durch umfangreichere Gedächtnisleistungen beansprucht.

Bei der quantitativen Auswertung der Tests kann jedes falsch geschriebene Wort als ein Fehler angesehen werden. Es gibt aber viele verschiedene Möglichkeiten, ein Wort falsch zu schreiben, auch können pro Wort mehrere Fehler auftreten. Solche Differenzierungen sind für eine qualitative Fehleranalyse zentral, welche wiederum unabdingbar für (förder-)diagnostisch relevante Schlüsse ist. Die Fehlertypologie sollte orthografietheoretisch fundiert sein und nicht auf problematischen Kausalannahmen beruhen, was beispielsweise bei Kategorien wie „Flüchtigkeitsfehler" der Fall ist.

Es liegen verschiedene Ansätze für solche Klassifikationen von Fehlern vor, etwa der Salzburger Lese- und Rechtschreibtest (SLRT) (vgl. Landerl, Wimmer & Moser, 1997), die Dortmunder Schriftkompetenzermittlung (DoSE) auf Grundlage der Dortmunder Rechtschreibfehler-Analyse (DoRA) (vgl. Löffler & Meyer-Schepers, 1992), die Hamburger Schreibprobe (HSP) (vgl. May, 2002), die Oldenburger Fehleranalyse (OLFA) (vgl. Thomé & Thomé, 2004) sowie die Aachener Förderdiagnostische Rechtschreibfehler-Analyse (AFRA) (vgl. Herné & Naumann, 2005). Einige dieser Verfahren sollen kurz näher vorgestellt werden.

Im Rahmen der IGLU-Studie 2001 wurde die Dortmunder Schriftkompetenzermittlung (DoSE) von Löffler und Meyer-Schepers eingesetzt, welche auf der Dortmunder

Rechtschreibfehler-Analyse (DoRA) (vgl. Löffler & Meyer-Schepers, 1992) beruht. Bei DoSE handelt es sich um einen Lückentext, der aus 19 Sätzen und 45 Testwörtern besteht. Die reine Diktatzeit wird auf 15 bis 25 Minuten angesetzt. Für die Auswertung können quantitative und qualitative Maße herangezogen werden. Bei der Beschreibung der rechtschreibbezogenen Kompetenzstufen wird zwischen elementarer und erweiterter Kompetenz in den Teilbereichen der Lautanalyse sowie der Grammatik differenziert.

Die Hamburger Schreibprobe (HSP) (vgl. May, 2002) ist ein Instrument der Diagnose orthografischer Kompetenz von der ersten bis zur neunten Jahrgangsstufe. Neben der Schreibung von Wörtern wird auch die von Graphemen erhoben. So können Schlüsse auf den Grad der Beherrschung der verschiedenen Rechtschreibstrategien (alphabetische, orthografische, morphematische und wortübergreifende Strategie)[3] gezogen werden. Als Testmaterial kommen Lückenwörter und Satzdiktate zum Einsatz. Die Fehlertypologie ist zwar einerseits weniger differenziert als bei anderen Ansätzen, andererseits ist die empirische Fundierung überzeugender.

Der orthografietheoretisch plausible Ansatz, auf der Basis von Häufigkeitsanalysen zwischen Mehrheits- und Minderheitsschreibungen beziehungsweise Basis- und Orthographemen zu unterscheiden, wird in der Oldenburger Fehleranalyse OLFA (vgl. Thomé & Thomé, 2004) und im Rahmen der Aachener Förderdiagnostischen Rechtschreibfehler-Analyse AFRA (vgl. Herné & Naumann, 2005) umgesetzt. OLFA ist mit 35 Kategorien deutlich umfangreicher als AFRA. Letztere versammelt auf vier „Ebenen" insgesamt 25 Kategorien, acht unter dem Etikett „Phonem-Graphem-Korrespondenz", sieben unter „Vokalquantität", sieben unter „Morphologie" und drei unter „Syntax". Ein solches Modell ist sehr voraussetzungsreich. Dabei geht es zum Beispiel um umstrittene Hypothesen über Art und Zahl von orthografischen „Prinzipien" – ist es sinnvoll, nicht nur ein phonematisches, morphematisches und syntaktisches, sondern auch ein silbisches und andere Prinzipien anzusetzen? Ferner geht es um Auseinandersetzungen in Bezug auf die Frage, inwiefern die Groß- und Kleinschreibung syntaktisch (über die Stellung im Satz) oder lexikalisch (über die Wortart) zu verstehen ist. Auf all das wird hier nicht weiter eingegangen (vgl. Herné & Naumann, 2005).

Die Tests zur Ermittlung der Rechtschreibkompetenz ermöglichen eine Diagnostik auf Individualebene, welche nicht nur ein globales Maß der Fähigkeit in diesem Bereich liefert, sondern auch im Sinne eines differenzierten Leistungsprofils Aussagen über spezifische Stärken und Schwächen gestattet. Allerdings können diese Testinstrumente mitunter bestimmten testdiagnostischen Qualitätsanforderungen nicht in vollem Ausmaß genügen. Zumeist werden für

[3] May (2002) unterscheidet eine orthografische von einer morphematischen Strategie. Zusätzlich macht er eine wortübergreifende Strategie geltend, die vor allem bei der Groß- und Kleinschreibung, der Getrennt- und Zusammenschreibung und der Zeichensetzung zum Tragen kommt.

die unterschiedenen Fehlerarten keine Skalenreliabilitäten und Ergebnisse zur Konstruktdimensionalität berichtet. Dies liegt jedoch sicherlich darin begründet, dass sie einer stärker fachdidaktischen und weniger psychometrischen Tradition folgen.

Zusammenfassend kann festgehalten werden, dass bereits vielfältige Instrumente für die Diagnose der Orthografiekompetenz vorliegen. Zum Teil erscheinen die Fehlertypologien aber als nicht differenziert genug, zum Teil steht die empirische Fundierung noch aus. Außerdem sind diese Instrumente nicht für den Zweck einer Operationalisierung der Bildungsstandards entwickelt worden. Es ist also wünschenswert, unter den in vielerlei Hinsicht beschränkenden Bedingungen eines Large-Scale-Assessments unter Rückgriff auf bewährte Verfahren Testinstrumente zu entwickeln, die auf einem orthografietheoretisch plausiblen Kompetenzmodell beruhen und ein reliables Globalmaß für die allgemeine orthografische Kompetenz von Schülerinnen und Schülern der dritten und vierten Jahrgangsstufe liefern. Eine weitergehende Differenzierung auf der Basis qualitativer Fehleranalysen unter Berücksichtigung von Basis- und Orthographemen wäre wünschenswert.

III.6 Kontextualisierung und Präzisierung der Fragestellungen

Standardbezogene Testverfahren können der Überprüfung von Kompetenzmodellen, dem Systemmonitoring, der Schulevaluation und – im Prinzip – auch der Individualdiagnostik dienen (vgl. Klieme et al., 2003). Bislang ist Individualdiagnostik aber nicht das Ziel großer Schulleistungsstudien, die im Zusammenhang mit der Evaluierung der Bildungsstandards durchgeführt werden. Dies begründet sich in erster Linie durch die Intention, Bildungsmonitoring auf Systemebene durchzuführen. Als Folge dieser Festlegung wird eine hinreichende Messgenauigkeit auf System-, aber nicht Individualebene angestrebt, sodass Aussagen über die Leistungsstände einzelner Schülerinnen und Schüler unpräzise bleiben. Pro Kompetenzbereich können bei geringer Testzeit nur wenige Aufgaben gestellt werden. Darüber hinaus werden den Schülerinnen und Schülern im Rahmen eines Multi-Matrix-Testdesigns unterschiedliche Aufgaben vorgelegt.[4]

Längerfristig könnte aber auch im Kontext großer Leistungsstudien und Vergleichsarbeiten verstärkt auf Individualdiagnostik abgezielt werden. Einen solchen Anspruch verfolgen bereits die Vergleichsarbeiten in der Grundschule (VERA 3). Für einen solchen Ansatz ist es jedoch zunächst unerlässlich, für ein orthografietheoretisch plausibles Kompetenzmodell

[4] In einem solchen Design bearbeitet jeder Teilnehmer nur eine kleine Menge der insgesamt im Test enthaltenen Items (vgl. Winkelmann & Böhme, 2009).

empirische Unterstützung bereitzustellen, damit dann auf Grundlage dieses Kompetenzmodells bestehende Lücken in den Bildungsstandards geschlossen und differenzierte Testinstrumente entwickelt werden können. Somit stehen zunächst die folgenden zwei Fragen im Zentrum:

1. Gestattet das entwickelte Testinstrument auf Wortebene eine Diagnostik der globalen Rechtschreibkompetenz, welche im Rahmen der Evaluierung der Bildungsstandards eine auf Schul- oder Klassenebene hinreichend reliable Aussage über den Leistungsstand von Schülerinnen und Schülern der dritten und vierten Jahrgangsstufen ermöglicht?
2. Kann darüber hinaus ein orthografietheoretisch fundiertes mehrdimensionales Modell der Rechtschreibkompetenz empirisch gestützt werden?

III.7 Anlage der Untersuchung

III.7.1 Testentwicklung

Es sollten also Testinstrumente bereitgestellt werden, die im Kontext von Large-Scale-Assessments eingesetzt werden können und trotz der besonderen und in vielerlei Hinsicht beschränkten Durchführungsbedingungen (z. B. stark begrenzte Testzeit, Multi-Matrix-Design) ein reliables Globalmaß für die orthografische Kompetenz von Schülerinnen und Schülern der dritten und vierten Jahrgangsstufe liefern. Darüber hinaus ist eine Differenzierung mithilfe qualitativer Fehleranalysen von Interesse.

Aus orthografietheoretischer Sicht wurde eine Diagnostik auf der orthografischen „Stufe" des Rechtschreiberwerbs angestrebt, gibt es doch zahlreiche Hinweise darauf, dass sehr viele Schülerinnen und Schüler in der dritten und vierten Klasse die alphabetische Strategie weitgehend beherrschen (vgl. May, 2002; vgl. Valtin et al., 2003). Außerdem wird verschiedentlich betont, dass insbesondere für die orthografische Phase des Rechtschreiberwerbs, „wie sie in den Klassen 3 und 4 erreicht sein dürfte, entsprechende Fehleranalysen auf empirisch breiter Basis [fehlen]" (Valtin et al., 2003, S. 229f.). So konstatiert auch Küttel: „Die Entwicklung der rechtschreiblichen Kompetenz ist für die ersten Erwerbsphasen ziemlich gut untersucht, aber für die Konsolidierungsabschnitte, die Ausdifferenzierung des orthografischen Könnens, mangelt es an aufschlussreichen Untersuchungen" (2003, S. 383).

Soll die Wahl der Testwörter und -wortformen nicht willkürlich sein und nicht allein auf einer Zufallsauswahl aus bundeslandspezifischen Grundwortschätzen beruhen, deren Herkunft oft ebenfalls recht dunkel ist, dann muss ein linguistisch plausibles System von Fehlerkategorien,

das ebenso plausibel auf ein Erwerbsmodell bezogen werden kann, Grundlage der Testentwicklung sein. Weiterhin sollte auf der Basis von Häufigkeitsanalysen zwischen Mehrheits- und Minderheitsschreibungen beziehungsweise Basis- und Orthographemen unterschieden werden. Deshalb gehen wir – ähnlich wie auch DoSE, die HSP, OLFA, AFRA – von einem Modell der Lupenstellen oder Fehlerkategorien aus. Bestimmte Stellen im Wort werden als Indikatoren für bestimmte orthografische Regularitäten verstanden beziehungsweise sind Fälle von Schwierigkeiten, die Lerner im Erwerbsprozess haben.

Diagnostische Kategorien

Unter Bezug auf die Aachener Förderdiagnostische Rechtschreibfehler-Analyse AFRA (vgl. Herné & Naumann, 2005) unterscheiden wir neun Fehlerkategorien und erheben zusätzlich die Auftretenshäufigkeit anderer, in den neun Kategorien nicht erfasster Fehler.

- *SG = spezielle Grapheme und Graphemverbindungen*: Selten sind zum Beispiel *x, v* (im Stamm), speziell sind die Verbindungen *qu, st, sp*.
- *VL+ = Vokallänge in der Mehrheit der Fälle:* In der weit überwiegenden Zahl der Fälle wird der lange betonte Vokal als einfacher Vokalbuchstabe geschrieben (*Hase, Vater*). Das lange betonte /i/ aber wird fast immer als <ie> verschriftet.
- *VL– = Vokallänge in der Minderheit der Fälle:* In bestimmten, deutlich weniger häufigen Fällen wird das Dehnungs-*h* oder das silbenanlautende *h* geschrieben (*Söh-ne* = Dehnungs-*h*, *Ru-he* = silbenanlautendes *h*). Fälle der Doppeltschreibung des Vokalbuchstabens fallen im Übrigen auch unter diese Kategorie, kommen hier aber nicht vor.
- *VK = Vokalkürze:* Folgt im Stamm auf einen kurzen betonten Vokal ein einzelner Konsonant, wird der Buchstabe für diesen Konsonanten (fast immer) doppelt hingeschrieben.
- *HM = häufige Morpheme:* Morpheme sind kleinste bedeutungstragende Einheiten. Dabei zählt beispielsweise auch eine Endung wie *ung* als Morphem, insofern sie anzeigt, dass es sich um ein Nomen handelt. Von Fall zu Fall ist allerdings fraglich, was als Morphem zählen kann, und darüber hinaus, ob eine Zuordnung zu HM oder zu einer anderen Kategorie, insbesondere zu KA (konsonantische Ableitung), sinnvoll ist.
- *MG = Morphemgrenze:* Treffen gleiche oder hinreichend ähnliche Laute an einer Morphemgrenze zusammen, verschmelzen sie im Mündlichen zu einem Laut. Im Geschriebenen werden solche Reduktionen meistens vermieden (bei *Fahrrad* z. B. hört

man nur ein /r/, man schreibt aber – nach Zerlegung in zwei Morpheme – zwei <r>). Zur Kategorie MG werden hier auch Fälle gezählt, bei denen Fugenelemente wie <s> zu schreiben sind.

- *VA = vokalische Ableitung:* Wörter mit demselben Stamm schreibt man weitgehend gleich. Beispielsweise lässt sich die *ä*-Schreibung bei *Bäcker* von der Schreibung *backen* ableiten.
- *KA = konsonantische Ableitung:* Am Ende eines Morphems oder in bestimmten Konsonantenverbindungen kann die Schreibung einiger Konsonanten nicht immer aus der Lautung erschlossen werden. In sehr vielen Fällen hilft hier das Verlängern (um eine Silbe). Schreibt man *lieblich* mit *ch* oder mit *g*? Die Verlängerung *lieb-li-cher* sorgt für Klarheit. (*-lich* könnte auch als häufiges Morphem – HM – verbucht werden. Führt die Verlängerungsprobe zum Ziel, wird im Folgenden immer „KA" und nicht „HM" gewählt.)
- *G/K = Groß- und Kleinschreibung:* Hier geht es vor allem um „Konkreta"; Abstrakta und Nominalisierungen sind selten (für die Beherrschung der Groß- und Kleinschreibung braucht man, soweit deklaratives Wissen im Spiel ist, Wissen um Wortarten und die Struktur von Satzgliedern).
- *AF = anderer Fehler:* Wenn im Rahmen des Tests nur die orthografische „Stufe" thematisiert werden soll, kann die Klassifikation der Fehler nicht erschöpfend sein. Die „alphabetischen" Fehler sowie Fehlschreibungen, die nicht in den neun oben genannten Kategorien klassifiziert werden, lassen sich in dieser Restklasse zusammenfassen.

Didaktische Rationale der Auswahl von Testwörtern

Für die Auswahl der konkreten Testwörter waren in Bezug auf die diagnostischen Kategorien vor allem folgende orthografietheoretische Maximen leitend:

- Angesichts dessen, dass einfache Phonem-Graphem-Korrespondenzen von den meisten (jedoch nicht von allen) Kindern im zweiten Schuljahr gemeistert werden, sollten vor allem solche Wörter vertreten sein, die bei AFRA als spezielle Grapheme (SG) und spezielle Verbindungen (SV) aufgeführt werden.
- Die Bereiche „Vokallänge" und „Vokalkürze" sind bis weit in die Sekundarstufe hinein fehlerträchtig. Es sollten daher diverse Wörter mit ungekennzeichnetem Langvokal (LV+) (z. B. V*o*gelfutter, k*a*putt) vorkommen und einige häufige und wenige seltene mit Dehnungs-h (z. B. L*eh*rerin, empf*eh*len). Auch beim langen <i> sollte die

Mehrheitsschreibung dominieren (z. B. vorsp*ie*len). Unbetonte Kurzvokale sollten nicht zu zahlreich sein, <tz> und <ck> als Sonderschreibungen für Doppelkonsonanten sollten vorkommen (z. B. fli*tz*en). An diversen Beispielen sollten die Schülerinnen und Schüler zeigen können, dass sie auf Morphemebene segmentieren können (z. B. warum *kennt*, aber *Wand*).[5]

- Auch anhand von Komposita, Derivationen und Flexionsformen sollten die Kinder demonstrieren, dass sie zur Morphemanalyse fähig sind. Es sollten häufige gebundene Morpheme wie *-ig, -lich, -heit* (z. B. unübertreff*lich*) und Präfixe wie *ver-* und *vor-* (z. B. *vor*spielen, *ver*staubten) vorkommen und viele Fälle vokalischer und konsonantischer Ableitung, allerdings vor allem in der Mehrheitsform (KA+ und VA+) (z. B. verstau*b*ten, Schlittschuhl*äu*fer).
- Darüber hinaus sollten großgeschriebene Konkreta dominieren und Substantive mit Artikelwort und typischem Suffix. Aber auch Abstrakta und wenige Substantivierungen (mit und ohne Artikelwort) sollten präsentiert werden.

Beschreibung der Testinstrumente

Aufgrund dieser Überlegungen wurden mithilfe des Orientierungswortschatzes von Naumann (1999) 80 Testwörter ausgewählt, die in insgesamt 40 Lückensätzen vorkamen. In jedem Satz waren also zwei Lücken zu ergänzen. Die Sätze wurden nach verschiedenen Geboten der Ausgewogenheit (z. B. Bekanntheit, Regelhaftigkeit, prognostizierte Schwierigkeit der Wörter sowie Auftreten der Fehlerkategorien) auf vier Testformen verteilt, sodass vier Testinstrumente mit jeweils zehn Lückensätzen beziehungsweise 20 Testwörtern entstanden. Für jede dieser Testformen wurde in Präpilotierungen eine Bearbeitungszeit von 20 Minuten erprobt, die sich in allen vier Fällen als großzügig erwies, sodass auch eher langsame Schreiber ohne Zeitdruck die fehlenden Wörter nach Diktat in der vorgesehenen Zeit ergänzen konnten.

Da die Bearbeitung der Lückensatzaufgaben während der Evaluierung der Bildungsstandards in einen umfangreicheren Test eingebettet war, in dem auch andere Kompetenzbereiche überprüft wurden, gab es für die Schülerinnen und Schüler zu Beginn des gesamten Tests eine ausführliche Einführung, wie die verschiedenen Aufgabenformate zu bearbeiten sind. Die Instruktion direkt vor Bearbeitung der Lückensatzdiktate lautete einheitlich: „*Die Sätze in dieser Aufgabe werden dir gleich vollständig vorgelesen. Lies bitte in deinem Heft mit! In den*

[5] Man kann sich auch an der Silbe orientieren und davon ausgehen, dass die Doppelschreibung des Konsonantenbuchstabens nicht die Vokalkürze, sondern das Silbengelenk – hier in der Stützform „kennen“ – anzeigt. Im Zweisilber „Wände“ gibt es hingegen kein Silbengelenk.

Sätzen im Heft fehlen immer zwei Wörter. Setze die fehlenden Wörter ein!" Daraufhin wurde der erste Satz vollständig vorgelesen. Anschließend wurde wiederholt, welches Wort in der ersten Lücke ergänzt werden sollte. Dann wurde das Wort wiederholt, welches in die zweite Lücke geschrieben werden sollte. Auf diese Weise wurde mit allen zehn Sätzen jeder Aufgabe verfahren. Auf Rückfrage wurden einzelne Testwörter vom Testleiter wiederholt.

III.7.2 Datengrundlage

Im Rahmen der Pilotierung der Bildungsstandardaufgaben für den Primarbereich im Frühjahr 2006 wurden die soeben beschriebenen vier Lückensatzaufgaben für die Diagnostik der Rechtschreibkompetenz eingesetzt.[6] Drei dieser Aufgaben wurden sowohl in der dritten als auch in der vierten Klasse vorgelegt, die Aufgabe „Lückensätze 4" nur in der vierten Klasse.

Insgesamt liegen für die quantitativen und qualitativen Fehleranalysen der vier Lückensatzaufgaben die Daten von 3 480 Schülerinnen und Schülern vor. Jungen und Mädchen und Schülerinnen und Schüler der dritten und vierten Jahrgangsstufe sind jeweils annähernd gleich vertreten. Die genauen Stichprobenumfänge und die Geschlechter- und Jahrgangsstufenverteilungen können der Tabelle III.1 entnommen werden.[7]

60 Items wurden von jeweils mehr als 1 000 Schülern bearbeitet, die 20 Items der Aufgabe „Lückensätze 4" nur von 549 Viertklässlern.

---

[6] Weitere Aufgaben, zum Beispiel zu den Kompetenzen „ableiten" und „korrigieren können", wurden zwar eingesetzt, werden an dieser Stelle aber nicht näher behandelt.

[7] Auf die Differenzierung zwischen Schülerinnen und Schülern mit und ohne Migrationshintergrund muss hier verzichtet werden, da für die meisten Items zu geringe Stichprobenumfänge für diese Subgruppen vorliegen. Aussagen zu dieser Differenzierung werden aufgrund der Daten der Normierungsstudie möglich sein.

Tabelle III.1: Stichprobenumfänge, Geschlechter- und Jahrgangsstufenverteilungen bei den vier Lückensatzaufgaben

Aufgaben-bezeichnung	Stichproben-umfang*	Geschlecht	Jahrgangsstufe
D078_Lückensätze 1	N = 1090	Jungen: 580 (53%) Mädchen: 510 (47%)	Drittklässler: 553 (51%) Viertklässler: 537 (49%)
D079_Lückensätze 2	N = 1050	Jungen: 520 (50%) Mädchen: 530 (50%)	Drittklässler: 532 (51%) Viertklässler: 518 (49%)
D080_Lückensätze 3	N = 1062	Jungen: 557 (52%) Mädchen: 505 (48%)	Drittklässler: 532 (50%) Viertklässler: 530 (50%)
D081_Lückensätze 4	N = 549	Jungen: 278 (51%) Mädchen: 271 (49%)	Drittklässler: --- Viertklässler: 549 (100%)

*Anmerkung: Die in der Tabelle angegebenen Stichprobenumfänge addieren sich nicht zu 3 480. Dies ist durch den Einsatz eines Multi-Matrix-Designs begründet. Einige Schülerinnen und Schüler haben sowohl die Aufgabe „Lückensätze 1“ als auch die Aufgabe „Lückensätze 3“ bearbeitet.

III.7.3 Statistische Analysen

Um deskriptive Befunde mitteilen zu können, wurden zunächst auf Wort- und Fehlerebene die prozentualen Lösungshäufigkeiten für die 20 Testwörter der Beispielaufgabe ermittelt. Für die empirische Prüfung der psychometrischen Struktur der Testinstrumente wurden verschiedene ein- und mehrdimensionale Modelle der probabilistischen Testtheorie (*Item Response Theory,* IRT) (vgl. Embretson & Reise, 2000) und Kognitive Diagnosemodelle (*Cognitive Diagnostic Models*, CDMs) (vgl. DiBello, Roussos & Stout, 2007; vgl. Rupp & Templin, 2008) eingesetzt. Der entscheidende Vorteil von CDM- und IRT-basierten Modellen ist die Möglichkeit, Schülerfähigkeiten und Itemschwierigkeiten im Rahmen eines Multi-Matrix-Designs auf einer gemeinsamen Skala abzubilden (vgl. De Boeck & Wilson, 2004) und messfehlerbereinigte Korrelationen zwischen verschiedenen Dimensionen auf latenter Ebene zu ermitteln. Wir verwendeten zur Berechnung der ein- und mehrdimensionalen Modelle der Rasch-Familie die Software *AcerConQuest* (vgl. Wu, Adams, Wilson & Haldane, 2007).

Der Unterschied zwischen beiden Modellklassen besteht in erster Linie darin, dass IRT-Modelle von kontinuierlichen Fähigkeitsskalen ausgehen, während bei CDMs diskrete Fähigkeitsskalen (meistens dichotome Fähigkeiten) als Ausgangspunkt einer multidimensionalen Klassifikation der Schülerinnen und Schüler verwendet werden. Sollte es möglich sein,

empirische Unterstützung für die angenommene Dimensionalität zu finden, sollten mit beiden Skalierungszugängen ähnliche Resultate erzielt werden.

Als weiterer Zugang der Dimensionalitätsprüfung wurden mithilfe der Software R (R Development Core Team, 2008) auf Fehlerebene für alle vier Aufgaben Hauptkomponentenanalysen (PCA) berechnet, wobei die Methode der Parallelanalyse (vgl. Bortz, 2005) zur Feststellung der Dimensionalität eingesetzt wurde. Nach Hattie (1985) lässt sich die Anzahl der Dimensionen anhand des Quotienten zweier aufeinanderfolgender Eigenwerte ableiten. Hierbei wird man sich für ein eindimensionales Modell entscheiden, wenn der Varianzanteil des ersten Faktors und das Varianzverhältnis aus erstem und zweitem Eigenwert möglichst groß ausfallen.

Da auf Fehlerebene dichotome Daten vorliegen, bildete eine Matrix tetrachorischer Korrelationen die Basis der PCA. Für solche dichotomen Daten wurde in der Literatur eine modifizierte Parallelanalyse vorgeschlagen, die als zu vergleichende Eigenwerte Daten aus einem eindimensionalen IRT-Modell simuliert (vgl. Finch & Monahan, 2008). Ähnlich wie bei der Hauptkomponentenanalyse kann so auch für dichotome Daten ein Parallelplot eingesetzt werden, der am Datensatz ermittelte Eigenwerte mit simulierten Eigenwerten unter Annahme unkorrelierter Variablen vergleicht. Ist dabei der zweite beobachtete Eigenwert kleiner als der zweite simulierte Eigenwert, so ist von Eindimensionalität auszugehen. Alternativ dazu kann man auch das Verhältnis aus zweitem beobachtetem und simuliertem Eigenwert einsetzen, das bei Eindimensionalität nicht wesentlich größer als eins ausfallen soll.

III.8 Ergebnisse

III.8.1 Deskriptive Befunde

Zur Beantwortung der Frage, inwieweit es gelungen ist, Testinstrumente zu entwickeln, die eine Beurteilung der globalen Rechtschreibfähigkeit von Schülerinnen und Schülern der dritten und vierten Jahrgangsstufe gestatten, betrachten wir zunächst die prozentualen Lösungshäufigkeiten auf Wortebene.[8] Diese sind in Tabelle III.2 dargestellt.

[8] Auf eine Darstellung der Fehlerdichte, wie zum Beispiel in der Berichtlegung zur IGLU-Studie 2001 verwendet (vgl. Valtin, Badel, Löffler, Meyer-Schepers & Vos, 2003) verzichten wir hier, da die Korrelation zwischen der Anzahl falsch geschriebener Wörter und der Fehlerdichte mit $r = .91$ sehr hoch ist und wir beide Informationen als redundant betrachten.

Deutlich wird, dass die drei beziehungsweise vier Testformen innerhalb der beiden Jahrgangsstufen jeweils auf sehr ähnlichen Schwierigkeitsniveaus messen. Für die Überprüfung der Frage, inwieweit die Ausdifferenzierung der Rechtschreibkompetenz auf der Ebene der orthografischen „Stufe" für Schülerinnen und Schüler ein angemessener diagnostischer Zugang ist oder ob zahlreiche Fehler an Stellen auftreten, die nicht durch die postulierten Lupenstellen abgedeckt werden konnten, wurde die Häufigkeit der Restkategorie des „anderen Fehlers" (AF) ermittelt.

Tabelle III.2: Prozentuale Lösungshäufigkeiten nach Aufgabe und Jahrgangsstufe

	Lösungshäufigkeit in %	
Aufgabe	**Klasse 3**	**Klasse 4**
D078	49.6	64.7
D079	45.0	64.2
D080	49.1	66.4
D081	---	67.0

Das Vorliegen eines anderen Fehlers wurde unabhängig von den diagnostischen Kategorien kodiert. Liegt die Häufigkeit eines „anderen Fehlers" für ein Wort beispielsweise bei zehn Prozent, bedeutet dies, dass zusätzlich zu den an den Lupenstellen vermerkten Fehlern bei diesem Wort von zehn Prozent aller Schülerinnen und Schüler (auch) an einer anderen Stelle des Wortes ein Fehler gemacht wurde. Über alle 80 Wörter hinweg bewegte sich die Auftretenshäufigkeit eines nicht durch die Lupenstellen abgedeckten Fehlers zwischen null und 58 Prozent. Kumulativ betrachtet lag die Häufigkeit eines „anderen Fehlers" für 25 der 80 Wörter unter zehn Prozent, für 53 der 80 Wörter unter 20 Prozent, für 67 der 80 Wörter unter 30 Prozent und lediglich für elf der 80 Wörter lag die Auftretenshäufigkeit eines nicht kategorisierten Fehlers zwischen 30 und 44 Prozent. Die zwei verbleibenden Wörter wiesen mit 50 und 58 Prozent Auftretenshäufigkeit eines „anderen Fehlers" recht hohe Werte auf, die jedoch in beiden Fällen auf Probleme beim Diktat zurückzuführen waren. Die weit überwiegende Mehrzahl der innerhalb der 80 Wörter gemachten Fehler lässt sich demnach mithilfe der neun postulierten Kategorien beziehungsweise Lupenstellen erfassen, was auf die Angemessenheit der Diagnostik auf der orthografischen Stufe hindeutet.

III.8.2 Geschlechts- und Jahrgangsdifferenzen

Tabelle III.3: Prozentuale Lösungshäufigkeiten für die 20 Testwörter der Aufgabe „Lückensätze 2“ (D079) nach Jahrgangsstufe und Geschlecht

Testwort	Lösungshäufigkeiten			
	Klasse 3	Klasse 4	Jungen	Mädchen
Lehrerin	67.0	85.5	73.4	78.9
versteckt	60.8	79.7	70.7	69.6
Verkehr	32.9	64.0	47.0	49.4
gesperrt	10.5	35.5	21.0	24.7
Staubsauger	63.8	80.5	70.3	73.7
kaputt	24.2	52.4	36.5	39.7
Geburtstag	55.6	69.4	56.4	68.4
Geschenke	71.9	83.2	75.1	79.8
vorspielen	55.2	73.7	65.5	63.1
eigentlich	63.3	74.7	66.7	71.1
flitzten*	30.5	37.1	33.3	34.2
Schlittschuhläufer	22.2	40.4	25.7	36.6
empfehlen	20.7	43.3	31.6	32.1
Kartoffelsuppe	44.5	69.4	52.6	60.9
verstaubten	34.6	51.7	41.8	44.2
Vogelfutter	69.2	86.7	73.3	82.3
unübertrefflich	28.4	48.9	33.9	43.1
verabschiedete	23.7	50.8	33.0	41.0
Schlüsselloch	57.6	70.4	59.2	68.6
Wohnzimmer	63.8	86.4	69.7	80.1

*Anmerkung: Beim Diktieren des Wortes „flitzten“ traten Probleme auf. Einige Testleiter diktierten hier „flitzen“ statt „flitzten“, weshalb die Varianz zwischen Klassen auffällig groß ausfällt.

Da eine vollständige Veröffentlichung der Testwörter und -aufgaben nicht möglich ist,[9] wird im Folgenden beispielhaft die Aufgabe „Lückensätze 2“ (D079) herausgegriffen. Alle Analysen beziehen sich, wenn nicht anders angegeben, auf diese Beispielaufgabe. In Tabelle III.3 sind die prozentualen Lösungshäufigkeiten aller 20 Testwörter der Aufgabe „Lückensätze 2“ getrennt für die dritte und vierte Jahrgangsstufe sowie für Jungen und Mädchen dargestellt. Im Anhang findet sich eine Auflistung des Vorkommens dieser Testwörter im Orientierungswortschatz nach Naumann (1999) und in häufig eingesetzten standardisierten Testverfahren.

In den Abbildungen III.1 und III.2 finden sich grafische Veranschaulichungen der Verteilungen für die Jahrgangsstufen 3 und 4 (Abbildung III.1) sowie für Jungen und Mädchen (Abbildung III.2). Dargestellt wird jeweils als „Violinplot“ die Dichtefunktion der Verteilung, der Interquartilsbereich als dunkelgrauer Boxplot, der Median der Verteilung als weißes Dreieck innerhalb der Box, der Mittelwert der Verteilung als weißer Punkt innerhalb der Box sowie das Konfidenzintervall des Mittelwerts, in dem als Fehlerbalken vom Mittelwert +/– zweimal der Standardfehler des Mittelwerts abgetragen wurde.

Betrachtet man die Leistungsverteilungen für die dritte und vierte Jahrgangsstufe, so erkennt man, dass die vierte Jahrgangsstufe einen deutlichen Leistungsvorsprung hat, der im Fall der Beispielaufgabe „Lückensätze 2“ (D079) einer Effektstärke von $d = 0.77$ entspricht. Über alle drei Testformen, die in beiden Jahrgangsstufen eingesetzt wurden, ergibt sich im Mittel ein Effekt von $d = 0.70$.

[9] Es muss sichergestellt sein, dass sie erneut im Rahmen der Normierung der Bildungsstandards und ab dem Jahr 2011 für die geplanten Bundesländervergleiche (vgl. KMK, 2006) verwendet werden können.

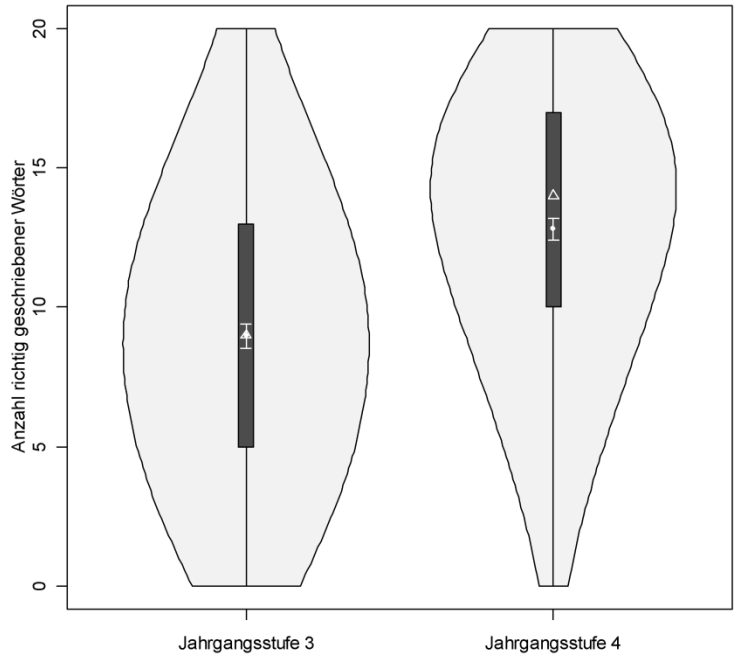

Abbildung III.1: Verteilung (Violinplots) für die Anzahl richtiger Wörter in der dritten und vierten Jahrgangsstufe für die Aufgabe „Lückensätze 2“ (D079)

Wie man in Abbildung III.1 erkennen kann, ist die Verteilung der vierten Jahrgangsstufe deutlich linksschief. Der Median, dargestellt als Dreieck, liegt oberhalb des Mittelwertes, abgebildet als Punkt. Dieser Befund weist auf leichte Deckeneffekte hin, was bedeutet, dass in der vierten Jahrgangsstufe im unteren Leistungsbereich bessere Differenzierungsmöglichkeiten bestehen als im oberen. Ein ähnlicher Befund zeigte sich im Rahmen der IGLU-Studie 2001, wo zur Überprüfung der orthografischen Kompetenz der Schülerinnen und Schüler der Rechtschreibtest DoSE (vgl. Löffler & Meyer-Schepers, 1992) eingesetzt wurde. Auch dort ergab sich eine leicht linksschiefe Verteilung (vgl. Valtin et al., 2003).

Beim Vergleich der Rechtschreibleistungen von Jungen und Mädchen, grafisch dargestellt in Abbildung III.2, zeigt sich für die überwiegende Zahl der Testwörter ein teils deutlicher Leistungsvorsprung der Mädchen. Im Mittel über alle 20 Wörter der Beispielaufgabe „Lückensätze 2“ (D079) ergibt sich eine Effektstärke von $d = 0.19$. Betrachtet man die Leistungsdifferenz zwischen Jungen und Mädchen über alle vier eingesetzten Testformen hinweg, so resultiert auch insgesamt ein sehr ähnlicher Effekt von $d = 0.22$. Für die Aufgabe „Lückensätze 2“ (D079) zeigten sich keine statistisch oder inhaltlich bedeutsamen Interaktionseffekte zwischen Geschlecht und Jahrgangsstufe.

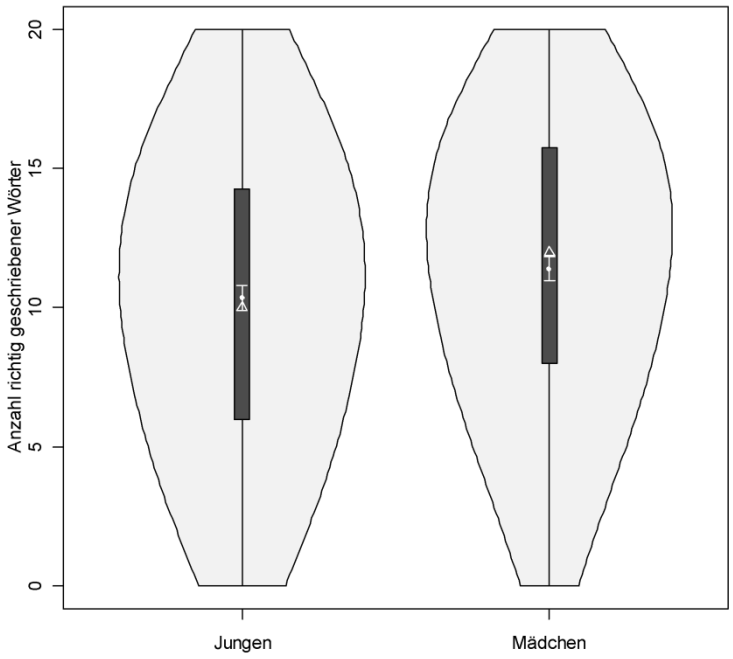

Abbildung III.2: Verteilung (Violinplots) für die Anzahl von Jungen und Mädchen richtig geschriebener Wörter für die Aufgabe „Lückensätze 2“ (D079)

III.8.3 Psychometrische Kennwerte und Dimensionsanalysen

Für die Klärung der Konstruktstruktur betrachten wir zunächst solche Skalierungen, bei denen von einer Eindimensionalität des Konstrukts ausgegangen wird und somit alle Items als Indikatoren eines homogenen Merkmals verstanden werden. Hier interessiert zum einen die Frage nach der Passung des Rasch-Modells, welche mithilfe der Infit-Maße für die einzelnen Items beantwortet werden kann. Zum anderen ist die Reliabilität der Messung von Interesse.

Auf Wortebene zeigen von den 20 in der Aufgabe „Lückensätze 2“ (D079) verwendeten Wörtern in einem eindimensionalen Rasch-Modell zwei Wörter einen unbefriedigenden Itemfit. Hierbei handelt es sich um die Wörter „eigentlich“ und „flitzten“. Auf Fehlerebene wird mithilfe von 56 Lupenstellen diagnostiziert. Von diesen 56 Items weisen in einem eindimensionalen Modell lediglich vier einen Infit von über 1.20 auf, der auf eine schlechte Passung mit dem Modell hindeutet.

Tabelle III.4 zeigt die in IRT-Skalierungen gewonnenen Reliabilitätsschätzungen sowohl auf Wortebene (Spalten 2 und 3) als auch auf Fehlerebene, also auf der Ebene der Lupenstellen (Spalten 4 und 5), jeweils getrennt für die dritte und vierte Jahrgangsstufe. Diese Reliabilitätsschätzungen folgen der Idee von Mislevy, Beaton, Kaplan und Sheehan (1992),

wonach Reliabilität im Kontext von IRT-basierten Skalierungsmethoden ein Maß für die Reduktion der Unsicherheit hinsichtlich der Ausprägung einer latenten Fähigkeit eines Schülers oder einer Schülerin durch den Messprozess darstellt (vgl. Adams, 2005). Angegeben sind Maße der EAP/PV-Reliabilität, welche nach Adams und Carstensen (2002) der Korrelation zwischen unabhängigen Ziehungen von Plausible Values entsprechen. Die Angabe von Grenzwerten für zufriedenstellende oder gute Reliabilitäten ist an dieser Stelle schwierig, da die genannten Reliabilitätsmaße designabhängig sind (vgl. Adams, 2005). Ein direkter Vergleich mit Reliabilitätsangaben der klassischen Testtheorie, wie etwa Cronbachs α ist nur sehr bedingt möglich.

Tabelle III.4: Reliabilitätsschätzungen aus eindimensionalen Skalierungen auf Wort- und Lupenstellenebene

	auf Wortebene		auf Fehlerebene	
	Klasse 3	Klasse 4	Klasse 3	Klasse 4
D078	.803	.825	.865	.832
D079	.898	.843	.937	.835
D080	.826	.836	.826	.831
D081	---	.774	---	.715

Um die weitergehenden Fragen zur Struktur der Rechtschreibkompetenz zu untersuchen, wurden mehrdimensionale Modelle unterschiedlicher Komplexität betrachtet. Zunächst wurde ein Modell gewählt, welches alle neun vorgestellten Fehlerkategorien als jeweils eigenständige Dimensionen beinhaltet. Nach und nach wurde die Komplexität dieses Modells und somit die Anzahl der voneinander unterschiedenen Dimensionen reduziert. Um diese Komplexitätsreduktion zu erreichen, wurden verschiedene Dimensionen, die im ursprünglich neundimensionalen Modell noch voneinander unterschieden worden waren, gebündelt. Somit konnten neben einem neundimensionalen auch ein sechs- und ein dreidimensionales Modell untersucht werden.

Das neundimensionale Modell umfasst die in Abschnitt III.7 vorgestellten neun Fehlerkategorien, die jetzt als Dimensionen verstanden werden: Spezielle Grapheme (SG), Vokallänge in der Mehrheitsschreibung (VL+), Vokallänge in der Minderheitenschreibung (VL–), Vokalkürze (VK), Häufige Morpheme (HM), Morphemgrenzen (MG), Vokalische Ableitungen (VA), Konsonantische Ableitungen (KA) und Groß- und Kleinschreibung (GK).

Da die Komplexität dieses orthografietheoretisch basierten Modells sehr hoch ist, wurde aufgrund theoretischer Überlegungen ein Modell geringerer Komplexität definiert. In diesem Modell wurden die beiden Dimensionen der Vokallänge in der Mehrheits- und der Minderheitenschreibung zu einer Dimension zusammengefasst, ebenso die beiden Kategorien der häufigen Morpheme sowie der Morphemgrenzen zu einer Dimension „Morphologie". Weiterhin wurden vokalische und konsonantische Ableitungen zu einer Dimension „Ableitung" zusammengelegt. Somit ergibt sich folgendes sechsdimensionale Modell: Spezielle Grapheme (SG), Vokallänge (VL+, VL–), Vokalkürze (VK), Morphologie (HM, MG), Ableitungen (VA, KA) und Groß- und Kleinschreibung (GK). Auf der Ebene der Lupenstellen wurden neun- sowie sechsdimensionale Rasch-Modelle berechnet. In den neun- und sechsdimensionalen Kognitiven Diagnosemodellen wurde mit Daten auf Wortebene gearbeitet.

Es zeigten sich jeweils für beide Modellvarianten massive Konvergenzprobleme bei der Schätzung der Modellparameter, da im Verhältnis zur Itemanzahl sehr viele Dimensionen zu schätzen sind und demzufolge einige Dimensionen nur durch wenige Items repräsentiert werden. In der Regel führte der gewählte Algorithmus (hier: *Marginal Maximum Likelihood*, MML) zu keiner eindeutigen Parameterschätzung. Es war also nicht möglich, die Daten an das (jeweilige) Modell anzupassen. Über die Konvergenzprobleme hinaus ergaben sich für die neun- und sechsdimensionalen Modelle schlechte Modellpassungen und nicht plausibel interpretierbare Korrelationen zwischen den Dimensionen. Beispielhaft seien wiederum Ergebnisse für die Aufgabe „Lückensätze 2" (D079) wiedergegeben, für die als Einzige eine Konvergenz der Modelle erreicht werden konnte.

Korrelationen auf latenter Ebene zwischen den neun postulierten Dimensionen werden in Tabelle III.5 zusammengefasst. Adams und Carstensen äußern hierzu:

> It is important to note that these latent correlations are unbiased estimates of the true correlation between the underlying latent variables. As such they are not attenuated by the unreliability of the measures and will generally be higher than the typical product moment correlations that have not been disattenuated for unreliability. (2002, S. 152)

Die gefundenen Zusammenhänge sind nicht durchgängig plausibel. Das gilt beispielsweise für die teilweise sehr hohen Korrelationen zwischen der Skala „Spezielle Grapheme" mit den anderen Dimensionen, da Schreibungen der ersten Kategorie hauptsächlich durch Abruf aus dem lexikalischen Speicher generiert werden, die Schreibungen der anderen Kategorien jedoch deutlich regelgeleiteter sind (bzw. sein sollten).

In der Literatur wird diskutiert, dass es einen Unterschied macht, ob mehrdimensionale Strukturen in einem komplexen gemeinsamen Modell geprüft werden oder ob die

Zusammenhänge der einzelnen Dimensionen in kleineren Teilmodellen, beispielsweise zweidimensional modelliert werden (vgl. Fieuws & Verbeke, 2006). Im vorliegenden Fall erweisen sich die Befunde zweidimensionaler Analysen, bei denen für jede Aufgabe immer paarweise jeweils die Items zweier diagnostischer Kategorien gemeinsam skaliert wurden, jedoch als ganz ähnlich problematisch wie komplexe Modellierungen. Die gefundenen Beziehungen sind über die Aufgaben hinweg nicht stabil.

Tabelle III.5: Reliabilitätsschätzungen und Korrelationen auf latenter Ebene aus einer neundimensionalen Skalierung der Aufgabe „Lückensätze 2“ (D079)

	SG	VL+	VL-	VK	HM	MG	VA	KA	GK
SG	.790								
VL+	.843	.890							
VL-	.788	.966	.876						
VK	.815	.960	.936	.878					
HM	.909	.917	.907	.896	.867				
MG	.779	.946	.910	.925	.864	.780			
VA	.841	.927	.932	.928	.930	.876	.840		
KA	.664	.740	.735	.680	.754	.670	.677	.674	
GK	.716	.731	.723	.758	.690	.675	.698	.679	.668

Anmerkung: Auf der Diagonalen sind grau unterlegt die EAP/PV-Reliabilitäten der Skalen, unterhalb der Diagonalen die korrelativen Zusammenhänge auf latenter Ebene dargestellt.

Daher wurde in einem nächsten Reduktionsschritt eine weitere Vereinfachung des Modells vorgenommen, dessen Ergebnis drei verbleibende Dimensionen sind. Hierbei wird die Kategorie der speziellen Grapheme beibehalten, da in ihr Rechtschreibphänomene zusammengefasst werden, die besondere Phonem-Graphem-Zuordnungen betreffen (z. B. *Hexe, quer*). Neben den speziellen Graphemen wird eine Dimension der Dehnung und Schärfung angenommen, die alle Phänomene der Vokallänge und -kürze umfasst. Schließlich wird die Dimension der Morphologie erweitert und beinhaltet nun nicht nur häufige Morpheme und Morphemgrenzen, sondern auch vokalische und konsonantische Ableitungen sowie die Groß- und Kleinschreibung. Die drei

Dimensionen dieses stark reduzierten Modells lauten somit: Spezielle Grapheme (SG), Dehnung und Schärfung (VL+, VL– sowie VK) und Morphologie (HM, MG, VA, KA sowie GK).

Nur 13 der insgesamt 227 Lupenstellen weisen in einem dreidimensionalen Rasch-Modell einen Infit auf, der auf eine mangelhafte Passung mit dem Modell hindeutet (Infit größer 1.20). Reliabilitätsschätzungen sowie Korrelationen auf latenter Ebene für das dreidimensionale Rasch-Modell sind in Tabelle III.6 zusammengefasst. Die Höhe der Korrelationen macht deutlich, dass eine klare analytische Trennung der spezifizierten Dimensionen nicht gelingt, vielmehr sprechen die Befunde dafür, von einem homogenen Maß der Rechtschreibkompetenz auszugehen. Allerdings ist ein Vergleich der verschiedenen Modelle, zum Beispiel durch Verwendung von Informationskriterien wie AIC oder CAIC (vgl. Burnham & Anderson, 2004), an dieser Stelle nicht möglich, da nicht konvergierte Modellschätzungen nicht als plausible Lösungen gewertet werden können. Die neundimensionalen Modelle können nicht gegen die Modelle geringerer Dimensionalität getestet werden.

Tabelle III.6: Reliabilitätsschätzungen und Korrelationen auf latenter Ebene aus einer gemeinsamen dreidimensionalen Skalierung aller Aufgaben

	SG	DS	MO
SG	.766		
DS	.884	.839	
MO	.884	.909	.799

Anmerkung: Auf der Diagonalen sind grau unterlegt die EAP/PV-Reliabilitäten der Skalen, unterhalb der Diagonalen die korrelativen Zusammenhänge auf latenter Ebene dargestellt.

Da dies insgesamt wenig zufriedenstellen kann, wurde als weiterer Zugang zur Dimensionalitätsprüfung eine Hauptkomponentenanalyse auf der Grundlage der tetrachorischen Korrelationen der Daten auf Fehlerebene für alle vier Testformen berechnet. Hierbei zeigte sich übereinstimmend, dass ein Hauptfaktor extrahiert werden kann, der zwischen 26 und 34 Prozent (im Mittel 30 Prozent) der Varianz aufklärt, was möglicherweise als gering anmuten könnte, für ein eindimensionales Modell aber als erwartbar einzuordnen ist (vgl. Swaminathan, Hambleton & Rogers, 2007). Der Quotient aus erstem und zweitem Eigenwert bewegt sich für alle vier Aufgaben zwischen 5.31 und 8.28 und liegt im Mittel über alle vier Aufgaben bei 6.89, was als recht deutlicher Hinweis auf Eindimensionalität interpretiert werden kann. Betrachtet man zusätzlich den zweiten beobachteten sowie den zweiten simulierten Eigenwert, so ergibt sich ein

mittlerer Quotient von 2.26 mit einem Minimum von 1.89 und einen Maximum von 3.01. Stets ist der zweite beobachtete Eigenwert größer als der zweite simulierte Eigenwert. Dies wiederum spricht eher nicht für Eindimensionalität. Auch durch diesen datenanalytischen Zugang wird auf Fehlerebene also keine eindeutig mehr-, sondern eher eine eindimensionale Lösung nahegelegt.

Es kann somit festgehalten werden, dass verschiedene analytische Zugänge (IRT, CDM) zur Strukturprüfung des Konstrukts der Orthografiekompetenz im Falle komplexer Modelle ähnliche (Konvergenz-)Probleme hervorrufen und dass niedrigdimensionale Lösungen plausibler sind als höherdimensionale. Das Verfahren der PCA bestätigt, dass näherungsweise mit einem eindimensionalen Modell zur Beschreibung der Rechtschreibkompetenz gearbeitet werden kann.

III.9 Diskussion

Die Befundlage lässt erkennen, dass es gelungen ist, Testinstrumente zu konstruieren, die unter den Bedingungen des Large-Scale-Assessments eine reliable Messung der globalen Rechtschreibfähigkeit von Schülerinnen und Schülern der dritten und vierten Jahrgangsstufe gestatten. Diese Testverfahren beziehen sich auf Rechtschreibphänomene, die für die orthografische Phase des Erwerbs relevant sind. Dass auf diese Weise tatsächlich die weit überwiegende Mehrzahl der rechtschreibschwierigen Stellen als solche erkannt und auf Lupenstellenebene erfasst werden konnte, lässt sich dadurch belegen, dass andere, nicht kategorisierte Fehler nur relativ selten auftreten. Solche Fehler können zum Beispiel auf unzutreffende Phonem-Graphem-Korrespondenzen zurückgeführt werden. Die Reliabilität der Testinstrumente wurde in eindimensionalen IRT-Skalierungen sowohl auf der Wortebene als auch auf der Ebene der Lupenstellen nachgewiesen. Die Reliabilitätsschätzungen und die Maße der Modellpassung belegen, dass die Testinstrumente eine Beurteilung der globalen Rechtschreibkompetenz von Schülerinnen und Schülern der dritten und vierten Jahrgangsstufe erlauben.

Darüber hinaus galt unser Interesse der Frage, inwieweit sich für ein orthografietheoretisch fundiertes neundimensionales Modell der Rechtschreibkompetenz empirische Belege finden lassen. Hier zeigten sich für höherkomplexe Modellvarianten keine befriedigenden Ergebnisse. Alle Analysen auf neun- und sechsdimensionaler Ebene ergaben übereinstimmend für verschiedene Analysemethoden, dass eine hinreichend reliable Messung jeweils homogener und voneinander separierbarer Dimensionen auf der Basis der verfügbaren Daten nicht möglich ist.

Dies kann verschiedene Gründe haben. So könnten Mängel bei der Operationalisierung von eigentlich zutreffend konzeptualisierten Konstruktdimensionen vermutet werden. Diese

Mängel könnten sich in einer insgesamt und auch pro Testform zu geringen Zahl an verwendeten Indikatoren für einzelne Dimensionen niederschlagen. Dieses Problem tritt zum Beispiel bei der Fehlerkategorie der vokalischen Ableitungen auf, für die insgesamt nur sechs und in einer Testform gar keine Indikatoren vorliegen. Ferner zeigt sich, dass die Varianzen der verschiedenen postulierten Dimensionen stark voneinander abweichen und oftmals sehr gering sind. Dieses Problem wirkt sich auch auf die mehrdimensionalen Skalierungen aus. Durch diese messmethodischen Erschwernisse können aber nicht die über alle Dimensionen und Testversionen hinweg problematischen Ergebnisse erklärt werden.

Ein weiterer möglicher Ansatzpunkt ist daher die Überlegung, dass mit der Fokussierung auf regelgeleitete Schreibungen eine Unterschätzung des Stellenwertes des orthografischen Speichers einhergeht. Es liegt also nahe zu vermuten, dass die Schülerinnen und Schüler gegen Ende der Grundschulzeit in einem höheren Maße als angenommen Wörter als Ganzheiten aus dem lexikalischen Speicher abrufen. In diesem Fall würden die Lupenstellen eines Wortes dadurch überlagert, dass das Wort als Ganzes erinnert wird (wenn auch nicht notwendigerweise in der konventionellen Schreibung). Die Fehler in einem Wort wären dann, anders als bislang angenommen, nicht unabhängig voneinander. Anders ausgedrückt: Werden „Ganzheiten“ aus dem lexikalischen Speicher abgerufen, dann ist die Bedingung der lokalen Unabhängigkeit der Fehlerlupenstellen innerhalb eines Wortes verletzt. Dieser Vermutung wurde nachgegangen, indem in die IRT-Modelle Testleteffekte aufgenommen wurden. Jedes Wort wurde als ein Testlet behandelt. Mittlere Testleteffekte pro Wort, ermittelt mithilfe der Q3-Statistik nach Yen (vgl. Yen, 1984, 1993), zeigen jedoch nur sehr geringe Effekte in der Größenordnung von unter .05. Somit kann auch dieser Ansatz nicht als generelle Erklärung herangezogen werden.

Wie lassen sich die Ergebnisse zur Konstruktdimensionalität, welche auf eine Eindimensionalität der Rechtschreibkompetenz (im Bereich der orthografischen „Stufe“) hindeuten, nun einordnen?

- Ein methodisch sehr ähnlicher Ansatz zur Dimensionalitätsprüfung der Rechtschreibkompetenz bei Grundschülern findet sich bei Voss, Blatt und Kowalski (2007). Dort wird die Rechtschreibkompetenz der Kinder breiter erfasst, es werden auch Aspekte der „Wortbildung“ und des „wortübergreifenden Prinzips“ einbezogen. Bei der Prüfung von fünfdimensionalen Modellen ergeben sich Korrelationen auf latenter Ebene zwischen .70 und .99. Der höchste Zusammenhang besteht zwischen den Dimensionen „Phonografisch/silbischer Kernbereich“ sowie „Morphologischer Kernbereich“ mit einer Korrelation von .99. Die Schreibungen im Kernbereich werden daher als eine homogene Fähigkeit interpretiert. Hieraus könnte sich der Schluss

ergeben, dass die von uns angestrebte Differenzierung innerhalb der orthografischen Stufe eine zu feine Auflösung anstrebt und eine weiter gefasste Diagnostik über mehrere Stufen des Rechtschreiberwerbs eher eine Bestimmung von Teilkompetenzen gestatten könnte.

- Auch im Kontext von Studien zur Legasthenie beziehungsweise zu besonderen Lese- und Rechtschreibschwierigkeiten wird keine Binnendifferenzierung innerhalb der orthografischen Stufe angestrebt. Vielmehr ist man bemüht, wenigstens zwei Gruppen von Kindern zu unterscheiden, und zwar einmal solche, „deren Schwierigkeiten primär darin bestehen, eine Phonemfolge zu analysieren und den Segmenten passende Grapheme zuzuordnen. Diese Form von Lese- und Rechtschreibschwierigkeiten sollte von einer Form abgegrenzt werden können, deren Problem primär im unzureichenden Aufbau orthografischen Wissens besteht“ (Klicpera, Gasteiger-Klicpera & Schabmann, 2003, S. 155).
- Auf der Basis einer (allerdings nicht repräsentativen) Stichprobe von knapp 300 Schülerinnen und Schülern aus zwei dritten, drei vierten, vier fünften und drei sechsten Klassen kam Scheele, die sich ebenfalls auf die AFRA-Typologie stützte, zu folgendem Ergebnis: „Aus dem Pseudolängsschnitt lassen sich kontinuierliche Fortschritte der Probanden in allen betrachteten morphologischen Regelbereichen ableiten“ (2006, S. 392). Und sie schließt: „Die einzelnen Erwerbsprozesse [...] lassen sich nicht gleichermaßen voneinander trennen wie die lautorientierte (alphabetische) Stufe von der lautergänzenden (orthografischen) Stufe [...]“ (ebd.).
- Dass es sinnvoll ist, von der Eindimensionalität der Rechtschreibkompetenz auszugehen, konnte im Primarbereich auch für andere Sprachen gezeigt werden (vgl. Notenboom & Reitsma, 2003).

Die hier vorgestellten Ergebnisse können so verstanden werden, dass gute Rechtschreiber in allen Problembereichen der orthografischen „Stufe“ gute Leistungen erzielen, wohingegen für eher schwache Rechtschreiber viele Rechtschreibbesonderheiten eine Schwierigkeit darstellen, ohne dass spezifische Stärken und Schwächen identifiziert werden können. Der Befund hoher positiver Zusammenhänge zwischen den neun postulierten Dimensionen impliziert somit, dass nicht von einem differenziellen Erwerb bestimmter Aspekte ausgegangen werden kann, sondern dass der Erwerb vielmehr kumulativ oder parallel breit über alle Teilbereiche hinweg erfolgt. Daraus zu schließen, dass eine thematische Differenzierung im Rechtschreibunterricht überflüssig ist, wäre allerdings abwegig.

Es sollte auch bedacht werden, dass alle hier vorgestellten Ergebnisse vor dem Hintergrund des interindividuellen Paradigmas interpretiert werden müssen und keine Aussagen über intraindividuelle Leistungsentwicklungen gestatten.

III.10 Literatur

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation*, *31*, 162-172.

Adams, R. J. & Carstensen, C. (2002). Scaling outcomes. In R. J. Adams & M. Wu (Eds.), *Pisa 2000 technical report* (pp. 149-162). Paris: OECD.

Brügelmann, H. & Brinkmann, E. (1994). Stufen des Schriftspracherwerbs und Ansätze zu seiner Förderung. In H. Brügelmann & S. Richter (Hrsg.), *Wie wir recht schreiben lernen* (S. 44-52). Konstanz: Libelle.

Burnham, K. P. & Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research*, *33*, 261-304.

De Boeck, P. & Wilson, M. (2004). *Explanatory item response models*. New York: Springer.

DiBello, L. V., Roussos, L. A. & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics: Psychometrics* (pp. 979-1030). Amsterdam: Elsevier.

Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler*. Berlin: Springer Verlag.

Eichler, W. (1986). Zu Uta Frith' Dreiphasenmodell des Lesen- (und Schreiben-)Lernens. Oder: Lassen sich verschiedene Modelle des Schriftspracherwerbs aufeinander beziehen und weiterentwickeln? In G. Augst (Hrsg.): *New trends in graphemics and orthography* (S. 234-247). Berlin: de Gruyter.

Eisenberg, P. (2005). Phonem und Graphem. In *Duden – Die Grammatik* (S. 19-94). Mannheim: Bibliographisches Institut.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Fieuws, S. & Verbeke, G. (2006). Pairwise fitting of mixed models for the joint modelling of multivariate longitudinal profiles. Biometrics. *Biometrics*, *62*, 424-431.

Finch, H. & Monahan, P. (2008). A bootstrap generalization of modified parallel analysis for IRT dimensionality assessment. *Applied Measurement in Education*, *21*, 119-140.

Frith, U. (1986). Psychologische Aspekte des orthographischen Wissens: Entwicklung und Entwicklungsstörung. In G. Augst (Hrsg.), *New Trends in graphemics and orthography* (S. 218-233). Berlin: de Gruyter.

Günther, K. B. (1986). Ein Stufenmodell der Entwicklung kindlicher Lese- und Schreibstrategien. In H. Brügelmann (Hrsg.), *ABC und Schriftsprache. Rätsel für Kinder, Lehrer und Forscher* (S. 32-54). Konstanz: Faude.

Hattie, J. (1985). Methodology Review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*, 139-164.

Herné, K.-L. (2003). Rechtschreibtests. In U. Bredel, H. Günther, P. Klotz, J. Ossner & G. Siebert-Ott (Hrsg.), *Didaktik der deutschen Sprache: Ein Handbuch* (S. 883-897). Paderborn: Ferdinand Schöningh.

Herné, K.-L. & Naumann, C. L. (2005). *Aachener Förderdiagnostische Rechtschreibfehler-Analyse (AFRA). Systematische Einführung in die Praxis der Fehleranalyse mit Auswertungshilfen zu insgesamt 33 standardisierten Testverfahren als Kopiervorlagen* (4. Auflage). Aachen: Alfa Zentaurus.

Klicpera, C., Gasteiger-Klicpera, B. & Schabmann, A. (2003). Rechtschreibschwierigkeiten. In U. Bredel, H. Günther, P. Klotz, J. Ossner & G. Siebert-Ott (Hrsg.), *Didaktik der deutschen Sprache: Ein Handbuch* (S. 405-419). Paderborn: Ferdinand Schöningh.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M. et al. (2003). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise.* Bonn: Bundesministerium für Bildung und Forschung.

KMK (2004). siehe Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2004).

KMK (2005a). siehe Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2005a).

KMK (2005b). siehe Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2005b).

KMK (2006). siehe Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2006).

Küttel, H. (2003). Entwicklung der grammatischen Rechtschreibkenntnisse. In U. Bredel, H. Günther, P. Klotz, J. Ossner & G. Siebert-Ott (Hrsg.), *Didaktik der deutschen Sprache: Ein Handbuch* (S. 380-391). Paderborn: Ferdinand Schöningh.

Landerl, K., Wimmer, H. & Moser, E. (1997). *SLRT: Salzburger Lese- und Rechtschreibtest. Verfahren zur Differentialdiagnose von Störungen des Lesens und Schreibens für die 1. bis 4. Schulstufe.* Bern: Hans Huber.

Lennox, C. & Siegel, L. S. (1998). Phonological and orthographic processes in good and poor spellers. In C. Hulme & R. M. Joshi (Eds.): *Reading and spelling: Development and disorders* (S. 395-404). Mahwah, NJ: Erlbaum.

Löffler, I. & Meyer-Schepers, U. (1992). *DoRA: Dortmunder Rechtschreibfehler-Analyse – Zur Ermittlung des Schriftsprachstatus rechtschreibschwacher Schüler*. Dortmund: ILT.

May, P. (2002). *HSP 1-9. Diagnose orthographischer Kompetenz. Zur Erfassung der grundlegenden Rechtschreibstrategien mit der Hamburger Schreibprobe. Manual*. Hamburg: Verlag für Pädagogische Medien.

Mislevy, R. J., Beaton, A. E., Kaplan, B. & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*, 133-161.

Naumann, C. L. (1999). *Orientierungswortschatz. Die wichtigsten Wörter und Regeln für die Rechtschreibung. Klassen 1 bis 6*. Weinheim: Beltz.

Notenboom, A. & Reitsma, P. (2003). Investigating the dimensions of spelling ability. *Educational and Psychological Measurement, 63*, 1039-1059.

Ossner, J. (1996). Silbifizierung und Orthographie des Deutschen. *Linguistische Berichte*, *165*, 349-400.

R Development Core Team (2008). *R: A language and environment for statistical computing*. Wien: R Foundation for Statistical Computing.

Röber-Siekmeyer, C. & Tophinke, C. (2002). Prosodisch orientierte Untersuchungen zur Wahrnehmung von Schärfungswörtern von Kindern am Schriftanfang. In C. Röber-Siekmeyer & C. Tophinke (Hrsg.), *Schärfungsschreibung im Fokus. Zur schriftlichen Repräsentation sprachlicher Strukturen im Spannungsfeld von Sprachwissenschaft und Didaktik* (S. 106-143). Baltmannsweiler: Schneider Hohengehren.

Rupp, A. A. & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, *8*, 219-262.

Scheele, V. (2006). *Entwicklung fortgeschrittener Rechtschreibfertigkeiten. Ein Beitrag zum Erwerb der „orthographischen" Strategien*. Frankfurt am Main: Peter Lang.

Scheerer-Neumann, G. (1986). Wortspezifisch: ja – Wortbild: nein. Ein letztes Lebewohl an die Wortbildtheorie. In H. Brügelmann (Hrsg.), *ABC und Schriftsprache: Rätsel für Kinder, Lehrer und Forscher* (S. 171-185). Konstanz: Faude.

Scheerer-Neumann, G. (1998). Schriftspracherwerb: „The State of the Art" aus psychologischer Sicht. In L. Huber, G. Kegel & A. Speck-Hamdan (Hrsg.), *Einblicke in den Schriftspracherwerb* (S. 31-46). Braunschweig: Westermann.

Scheerer-Neumann, G. (2003). Rechtschreibschwäche im Kontext der Entwicklung. In I. M. Naegele & R. Valtin (Hrsg.), *LRS – Legasthenie in den Klassen 1–10. Handbuch der Lese-Rechtschreibschwierigkeiten.* Band 1: Grundlagen und Grundsätze der Lese-Rechtschreib-Förderung (6. Auflage) (S. 45-65). Weinheim: Beltz.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2004). *Bildungsstandards im Fach Deutsch für den Mittleren Schulabschluss. Beschluss vom 04. 12. 2003*. München: Wolters Kluwer.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2005a). *Bildungsstandards im Fach Deutsch für den Primarbereich (Jahrgangsstufe 4). Beschluss vom 15.10.2004.* München: Wolters Kluwer.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2005b). *Bildungsstandards im Fach Deutsch für den Hauptschulabschluss. Beschluss vom 15.10.2004.* München: Wolters Kluwer.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2006). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring. Beschluss vom 02. 06. 2006*. München: Wolters Kluwer.

Swaminathan, H., Hambleton, R. K. & Rogers, H. J. (2007). Assessing the fit of item response theory models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics,* Vol. 26: Psychometrics (pp. 683-718). Amsterdam: Elsevier.

Thomé, G. (2003). Entwicklung der basalen Rechtschreibkenntnisse. In U. Bredel, H. Günther, P. Klotz, J. Ossner & G. Siebert-Ott (Hrsg.), *Didaktik der deutschen Sprache: Ein Handbuch* (S. 369-379). Paderborn: Ferdinand Schöningh

Thomé, G. & Thomé, D. (2004). *OLFA – Oldenburger Fehleranalyse.* Oldenburg: Institut für sprachliche Bildung.

Valtin, R. (1988). Schriftspracherwerb als Entwicklungsprozess. *Grundschule*, *12*, 12-16.

Valtin, R., Badel, I., Löffler, I., Meyer-Schepers, U. & Voss, A. (2003). Orthographische Kompetenzen von Schülerinnen und Schülern der vierten Klasse. In W. Bos, E. M. Lamkes, M. Prenzel, G. Walther & R. Valtin (Hrsg.), *Erste Ergebnisse aus IGLU: Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich* (S. 227-264). Münster: Waxmann.

Voss, A., Blatt, I. & Kowalski, K. (2007). Zur Erfassung orthographischer Kompetenz in IGLU 2006. Dargestellt an einem sprachsystematischen Test auf Grundlage von Daten aus der IGLU- Voruntersuchung. *Didaktik Deutsch*, *23*, 15-32.

Winkelmann, H. & Böhme, K. (2009). Anlage und Durchführung der Pilotierung der Bildungsstandards. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik* (S. 31-41). Weinheim: Beltz.

Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. A. (2007). *ACER ConQuest. Version 2.0.*. Camberwell, Victoria: ACER Press (Australian Council for Educational Research).

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the Three-Parameter Logistic Model. *Applied Psychological Measurement*, *8*, 125-145

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187-213.

3.4 Kompetenzbereich III – Lesen

3.4.1 Der Textbegriff

Der Kompetenzbereich III der Bildungsstandards für das Fach Deutsch im Primarbereich (KMK, 2005a) ist überschrieben mit: Lesen – mit Texten und Medien umgehen. Aber nicht nur hier spielt der Textbegriff eine Rolle, auch im Bereich Zuhören wird vom Verstehen auditiver Texte gesprochen. Daher scheint es angezeigt, den Textbegriff noch einmal näher zu betrachten.

Im Kontext des Large-Scale-Assessments wird im Rahmen der PISA-Studie ein sehr weit gefasster Textbegriff etabliert. Dieser bezieht neben kontinuierlichen, also fortlaufend geschriebenen Texten auch nicht-kontinuierliche Texte, also bildhafte Darstellungen wie beispielsweise Diagramme, Karten, Tabellen oder Graphiken mit ein (vgl. Artelt, Stanat, Schneider & Schiefele, 2001).

In PIRLS beziehungsweise IGLU (Bos et al, 2007), welche sich als internationale Schulleistungsstudie der Ermittlung der Lesekompetenz von Schülerinnen und Schülern im Primarbereich widmet, werden bislang ausschließlich kontinuierliche literarische und Sachtexte eingesetzt. Hiervon abweichend werden bei der Überprüfung der Erreichung der Bildungsstandards im Kompetenzbereich Lesen auch altersgerechte diskontinuierliche Sachtexte, also Tabellen und Karten, beispielsweise in Form von Stunden- und Lageplänen als Textstimuli verwendet.[1]

In Anlehnung an den in PISA verwendeten Textbegriff wird also auch für die Operationalisierung der Bildungsstandards im Kompetenzbereich Lesen auf literarische Texte sowie kontinuierliche und diskontinuierliche Sachtexte zurückgegriffen. Dieses Verständnis des Textbegriffs erstreckt sich auf die Operationalisierung der Bildungsstandards im Fach Deutsch für den Primarbereich (KMK, 2005a) sowie für die Sekundarstufe I (KMK, 2004, 2005b).

Bei der gegenwärtig stattfindenden Entwicklung der Bildungsstandards für die Allgemeine Hochschulreife Deutsch wird in Anlehnung an die Einheitlichen Prüfungsanforderungen in der Abiturprüfung (EPA) Deutsch (KMK, 1989) ebenfalls ein weiter Textbegriff verwendet. In den EPA heißt es: „Der angenommene weite Textbegriff schließt literarische Texte, pragmatische Texte sowie Medienprodukte als Gestalt-Gehalt-Einheiten ein“ (KMK, 1989, S. 5). Hierbei

[1] Geringe Unterschiede in den Ergebnissen der beiden Grundschulstudien könnten daher auch auf diesen partiellen Mangel an Konstruktinvarianz zurückzuführen sein (vgl. Pietsch, Böhme, Robitzsch & Stubbe, 2009).

werden als literarische Textsorten beispielhaft Lyrik, Romane, Erzählungen und Dramen benannt (vgl. KMK, 1989, S. 6). Als relevante pragmatische Textsorten werden Essays, Reden, Abhandlungen, Leitartikel, Glossen und Predigten angegeben (KMK, 1989, S. 6). Als Medienprodukte gelten unter anderem Literaturverfilmungen, Hörspielsequenzen, Interviews oder Filmtrailer (KMK, 1989, S. 6). „Weit" ist der Textbegriff in den EPA Deutsch also in erster Linie durch die Einbeziehung verschiedenster auditiver und audio-visueller Medienprodukte, weniger hinsichtlich der Berücksichtigung von nicht-kontinuierlichen Texten.

Aus Perspektive der Textlinguistik wird das in den EPA etablierte und in PISA vorherrschende Verständnis des Textbegriffs als sehr – und eventuell zu – weit gefasst und in diesem Sinne als ungeeignet eingeschätzt, da es eine klare Abgrenzung des (nicht-kontinuierlichen) Textes von anderen Zeichenformen unmöglich mache (Busch & Stenschke, 2008). Um den Textbegriff zu präzisieren, werden daher Textualitätskriterien herangezogen, welche die Grundmerkmale eines Textes verdeutlichen sollen. In einem frühen, häufig zitierten Beitrag von De Beaugrande und Dressler (1981, S. 3) werden die folgenden sieben Kriterien der Textualität benannt: Kohäsion, Kohärenz, Intentionalität, Akzeptabilität, Informativität, Situationalität und Intertextualität. In der Textlinguistik wurden diese Kriterien häufig aufgegriffen, unterschiedlich gewichtet und weiterentwickelt. So charakterisieren beispielsweise Busch und Stenschke (2008, S. 229) einen Text in ihrem aktuellen Lehrwerk durch die folgenden sechs Kriterien:

1. Sprachlichkeit,
2. Schriftlichkeit,
3. Kohäsion (strukturell-grammatischer Zusammenhang),
4. Kohärenz (inhaltlich-thematischer Zusammenhang),
5. Funktionalität und
6. Sortenhaftigkeit.

Demnach ist ein Text im Sinne der Textlinguistik eine sprachliche und keine ikonische, also bildliche Einheit. Der kompetente Umgang mit tabellarischen oder kartographischen Darstellungen, Fließdiagrammen etc. würde demnach nicht unter das verstehende „Lesen" von Texten fallen. Neben der Sprachlichkeit ist auch die Schriftlichkeit ein konstitutives Textmerkmal. Allerdings wird auch innerhalb der Textlinguistik kontrovers diskutiert, ob der Textbegriff mündlichen und schriftlichen Sprachgebrauch gleichermaßen einbeziehen sollte. In der Deutschdidaktik vertritt beispielsweise Ehlich den Standpunkt, dass Texte nicht dadurch gekennzeichnet sind, dass sie schriftlich fixiert wurden, sondern dadurch, dass sie produziert

werden, um zu überdauern (vgl. Ehlich, 1983). In diesem Sinne kann es somit auch mündliche Texte geben. Auch hier wird im Rahmen der Überprüfung der Bildungsstandards für das Fach Deutsch ein weites Verständnis vertreten, da auch auditive Stimuli für die Leistungsmessung im Bereich Zuhören als Hörtexte bezeichnet werden.

In einer aktuellen Auseinandersetzung mit dem literaturwissenschaftlichen und linguistischen Textbegriff und seiner Wandlung schreibt Peyer (2010):

> Nicht mehr (nur) die sprachliche Einheit und ihre Struktur interessieren, sondern auch die Wissensbasis und die kognitiven Prozesse, aufgrund derer die Beteiligten Texte verstehen und produzieren. Texte existieren so gesehen nicht auf einer abstrakten Ebene des sprachlichen Systems, […], sondern sie müssen in ihrer pragmatischen und kognitiven Situiertheit untersucht werden […]. (S. 253)

Diese Sichtweise des Textbegriffs ist mit aktuellen Bestimmungen von Lesekompetenz und Leseverstehen sehr gut vereinbar, da auch in diesem Kontext das Augenmerk auf die kognitive Informationsverarbeitung und die aktive Konstruktion der Textbedeutung gerichtet wird.

Da in dem nachfolgenden empirischen Beitrag „Methodische Aspekte der Erfassung der Lesekompetenz“ keine vertiefte Auseinandersetzung mit dem Konstrukt der Lesekompetenz erfolgt, werden Fragen der Konstruktdefinition sowie der Operationalisierbarkeit der Standards im Kompetenzbereich III vorab in den folgenden zwei Abschnitten behandelt.

3.4.2 Lesekompetenz und kognitionspsychologische Grundlagen des Leseverstehens[2]

Lesekompetenz als die Fähigkeit, kontinuierliche, diskontinuierliche und multimediale Texte zu verstehen, ist eine Schlüsselqualifikation, die für eine erfolgreiche Lebensführung im gesellschaftlichen Kontext unerlässlich ist. Sie ist nicht nur eine zentrale Bedingung für den Wissenserwerb und für das lebenslange Lernen, sondern auch für die Teilnahme an der Kommunikation über gesellschaftlich relevante Themen (vgl. Bos et al., 2003; Bundesministerium für Bildung und Forschung [BMBF], 2007).

---

[2] Der Abschnitt „Lesekompetenz und kognitionspsychologische Grundlagen des Leseverstehens“ basiert teilweise auf Textpassagen des Beitrags von Bremerich-Vos, A. & Böhme, K. (2009a). Lesekompetenzdiagnostik – die Entwicklung eines Kompetenzstufenmodells für den Bereich Lesen. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 219-249). Weinheim: Beltz.

In der Tradition des *Literacy*-Konzepts ist mit Lesekompetenz eine Fähigkeit gemeint, die „erforderlich ist für die Bewältigung der charakteristischen Kommunikations- und Handlungsanforderungen, denen ein […] Gesellschaftsteilnehmer in seinem Alltag und Beruf begegnet“ (Hurrelmann, 2007, S. 21).

Vor allem im Rahmen von Lesesozialisationsforschung und Literaturdidaktik werden darüber hinaus Aspekte von Lesekompetenz erörtert, die im *Literacy*-Konzept nicht oder nur am Rande berücksichtigt sind. Unter Bezug auf bildungstheoretische Diskurse steht daher nicht nur die instrumentelle Funktion des Lesens beispielsweise als Mittel für Zwecke des Lernens im Zentrum der Aufmerksamkeit, es wird ebenso der Beitrag des Lesens zur Persönlichkeitsbildung thematisiert. So setzt Hurrelmann (2007, S. 22f.) Lesen mit der Herausbildung von ästhetischer und sprachlicher Sensibilität, Moralentwicklung und Empathiefähigkeit, Fremdverstehen und Teilhabe am kulturellen Gedächtnis in Verbindung. Neben kognitiven und motivationalen Aspekten werden somit auch die emotionale Beteiligung und ferner die Bereitschaft und Fähigkeit zur Anschlusskommunikation, also zum Diskurs über das Gelesene, als Facetten der Lesekompetenz angesehen.

Im Rahmen großer Schulleistungsstudien wie PISA (Baumert et al., 2001) oder PIRLS beziehungsweise IGLU (Bos et al., 2007) wird allerdings zumeist nicht das hier angedeutete Verständnis von Lesekompetenz zugrunde gelegt, sondern einem pragmatisch-funktionalen Begriff der *Reading Literacy* gefolgt, wobei in der testdiagnostischen Überprüfung der Lesekompetenz die kognitive Dimension, also das Lese- oder Textverstehen, dominiert.

Psychologische Theorien des Textverstehens beruhen auf der zentralen Annahme, dass der Verstehensprozess als Konstruktionsleistung zu fassen ist. So definieren auch Artelt und Kollegen im Rahmen der PISA-Studie des Jahres 2000:

> Lesen ist keine passive Rezeption dessen, was im jeweiligen Text an Information enthalten ist, sondern aktive (Re-)Konstruktion der Textbedeutung. Die im Text enthaltenen Aussagen werden aktiv mit dem Vor-, Welt- und Sprachwissen des Lesers verbunden. Die Auseinandersetzung mit dem Text lässt sich als ein Akt der Bedeutungsgenerierung verstehen, bei dem das Vorwissen der Leser und die objektive Textvorgabe interagieren. (Artelt et al., 2001, S. 70 f.)

Von Seiten der Deutschdidaktik wird in diesem Zusammenhang regelmäßig darauf hingewiesen, dass beim Leseverstehen nicht von „Bedeutungsentnahme“, sondern davon gesprochen werden sollte, dass dem Text Informationen abgewonnen werden (vgl. Bremerich-Vos & Wieler, 2003; Grzesik, 2005).

Der Prozess des Textverstehens kann als komplexe Interaktion textgeleiteter und (vor-)wissensgeleiteter kognitiver Informationsverarbeitungsprozesse verstanden werden. Hierarchieniedrig, das heißt bei geübten, fluenten Lesern automatisiert, sind Prozesse auf subsemantischer Ebene wie die Identifikation von Graphemen und Silben sowie ihre Zuordnung zu lautlichen Einheiten, aber auch das Dekodieren, das heißt das Erfassen von Wortbedeutungen, darüber hinaus die (Re-)Konstruktion der Satzbedeutung. Häufig wird zwischen verschiedenen Formen der mentalen Repräsentation von Bedeutung unterschieden. Neben einer an der sprachlichen Oberfläche orientierten Repräsentation postuliert man eine propositionale Form sowie eine Repräsentation als mentales Modell beziehungsweise Situationsmodell.

Propositionen als Relationen von Prädikaten und Argumenten sind im Verständnis von Kintsch (1998) nicht-sprachliche Einheiten:

> Psychologists studying language processing need a better representation of meaning than is provided by the words and sentences of the language itself, that is, a representation that more directly reflects the semantic relations that are crucial for how people understand, remember, and think with language. (S. 49)

In der psychologischen Textverstehensforschung wurden intensive Bemühungen unternommen, Propositionen als mentale Repräsentationsform nachzuweisen (vgl. bspw. Kintsch, 1998). Im deutschsprachigen Raum gibt es eine kritische Auseinandersetzung mit der Rolle von Propositionen für das Textverstehen und einer auf dieser Annahme basierenden Testkonstruktion (vgl. bspw. Bremerich-Vos & Wieler, 2003; Grzesik, 2003). Grzesik (2003) betont, dass die Brauchbarkeit von Propositionen für die Messung von Verstehen begrenzt ist und äußert: „Es gibt keinen systematischen Zusammenhang zwischen dem Forschungsansatz […] von Kintsch mit den Aufgaben des internationalen PISA-Tests“ (Grzesik, 2003, S. 158).

Wenn Sätze auf lokaler Ebene miteinander verknüpft werden, können Kohäsionsmittel wie Konjunktionen, Adverbien und Präpositionen eine Rolle spielen. Fehlen sie, ergeben sich semantische Beziehungen aber auch als Resultate von intendierten Inferenzen, von textnahen Schlüssen. So kann zum Beispiel zwischen benachbarten Sätzen eine kausale Relation hergestellt werden, die im Text selbst nicht zu finden ist. Wird nicht nur lokal, sondern auf Textebene Kohärenz hergestellt, sind Operationen wie Auslassen und Auswählen, Verallgemeinern und Integrieren beteiligt. Im Rahmen von elaborativen Prozessen wird das, was sich als Textbasis verstehen lässt, in vielfältiger Weise mit dem Vorwissen, den Zielsetzungen und Erwartungen des Lesenden verknüpft. Die resultierende kognitive Repräsentation des Textinhalts wird als mentales oder Situationsmodell bezeichnet. Solche Modelle können allerdings nicht mit bildhaften Vorstellungen gleichgesetzt werden (vgl. Schiefele, 1996; Schnotz, 1994). „Kennzeichnend für ein

mentales Modell ist lediglich, dass es bestimmte Eigenschaften besitzt, die den zu repräsentierenden Eigenschaften des Originals funktional analog sind“ (Schnotz, 1994, S. 159).

3.4.3 Die Bildungsstandards im Kompetenzbereich Lesen und Möglichkeiten ihrer Testung

Die Standards im Kompetenzbereich „Lesen – mit Texten und Medien umgehen“ gruppieren sich in die vier Kategorien:

- über Lesefähigkeiten verfügen,
- über Leseerfahrungen verfügen,
- Texte erschließen,
- Texte präsentieren (vgl. KMK, 2005a, S. 11f.).

Der Kompetenzbereich Lesen umfasst in diesen vier Gruppen insgesamt 26 Standards, die allerdings nicht distinktiv sind. Verschiedene Standards sind gar nicht beziehungsweise nicht in ökonomisch vertretbarer Weise in Testaufgaben umsetzbar. Dies soll anhand einiger Beispiele veranschaulicht werden.

So ist etwa die Mitwirkung bei Lesungen und Aufführungen (subsumiert unter *Texte präsentieren*) nicht in einem standardisierten Testverfahren im Rahmen eines Large-Scale-Assessments testbar. Ebenso sind die Standards „Kinderliteratur kennen: Werke, Autoren und Autorinnen, Figuren, Handlungen“ sowie „über Leseerfahrungen verfügen“ im Rahmen eines Large-Scale-Assessments nicht operationalisierbar, da unter anderem die Testfairness beeinträchtigt wäre: Teilgruppen der Schülerinnen und Schüler, die mit den ausgewählten Autoren beziehungsweise Werken nicht vertraut sind, wären benachteiligt. Im Fokus der Testung stünde ferner das (Vor-)Wissen und nicht die Lesekompetenz der Kinder.

Eine Operationalisierung des Substandards „handelnd mit Texten umgehen: z. B. illustrieren, inszenieren, umgestalten, collagieren“ wäre zwar grundsätzlich möglich, fraglich wäre hier aber die Entwicklung objektiver Auswertungskriterien. Als nicht testbar erscheint ferner der auf elaborative Prozesse zielende Standard „lebendige Vorstellungen beim Lesen und Hören literarischer Texte entwickeln“. Denn wer mag entscheiden, wann Vorstellungen als lebendig zu werten sind, und wie valide kann eine Testung sein, wenn Kinder bildliche Vorstellungen in verbaler Form, womöglich sogar schriftlich, präsentieren müssen?

Für die Umsetzung in Testaufgaben kommen somit lediglich solche Standards in Betracht, für die präzise Konstruktdefinitionen vorliegen oder auf der Basis vorhandener

Theorien erarbeitet werden können und die eine Operationalisierung gemäß den Gütekriterien für standardisierte Testungen gestatten (vgl. Granzer, Böhme & Köller, 2008). Hierbei handelt es sich im Wesentlichen um Standards der Kategorie „Texte erschließen"; hinzu kommt der unter dem Etikett „über Lesefähigkeiten verfügen" gefasste, auf die Konstruktion eines Situationsmodells zielende zentrale Substandard „altersgemäße Texte sinnverstehend lesen".

Dem Postulat, dass die Bildungsstandards als Bestimmung angestrebter Kompetenzen in (Test-)Aufgaben umsetzbar sein sollen, um „prinzipiell mithilfe von Testverfahren" erfasst werden zu können (Klieme et al., 2007, S. 9), genügen die Standards für das Fach Deutsch im Primarbereich demnach allenfalls partiell (vgl. Abschnitte 2.2.2 sowie 4.2).

IV. Beitrag 4: Methodische Aspekte der Erfassung der Lesekompetenz

Autoren:
Katrin Böhme / Alexander Robitzsch

Erschienen in:
D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.) (2009a). *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 250-289). Weinheim: Beltz.

In den letzten Jahren wurden auch im deutschsprachigen Raum zahlreiche große Schulleistungsstudien durchgeführt. In diesem Forschungsfeld rücken zunehmend methodische Details und Besonderheiten ins Blickfeld, die bislang kaum thematisiert wurden. Dies betrifft beispielsweise die Fragen nach methodischen Ansätzen zur Bestimmung der Konstruktdimensionalität oder Aspekte des differenziellen Itemfunktionierens in leistungsheterogenen Gruppen. Beide Themen werden im vorliegenden Kapitel für den Bereich der Lesekompetenz aufgegriffen und anhand empirischer Daten besprochen.

Dimensionsanalysen auf der Basis parametrischer und nichtparametrischer Verfahren ergeben keine eindeutigen Hinweise auf eine mehrdimensionale Struktur der Lesekompetenz im Primarbereich, die einer Replikation der für die Sekundarstufe im Rahmen der PISA-Studie etablierten Dimensionalität (vgl. Artelt & Schlagmüller, 2004) entsprechen würden.

In Bezug auf die Verlinkung der untersuchten Jahrgangsstufen drei und vier mit gemeinsamen Items ergab sich, dass eine hinreichend dichte Verlinkung unbedingte Voraussetzung für die valide Bestimmung des Leistungsunterschieds zwischen diesen Gruppen ist. Die Wahl der Analysemethode zur Verlinkung (getrennte Skalierung mit anschließendem Equating vs. gemeinsame Skalierung mit Hintergrundmodell) ist hierbei eher sekundär, solange eine ausreichende Verbindung der betreffenden Gruppen mittels überlappender Itemmengen im Design hergestellt wurde.

IV.1 Einführung und Überblick

Bei der Entwicklung des von Bremerich-Vos und Böhme (2009) vorgestellten Kompetenzstufenmodells für den Bereich Lesen wurden in unserer Forschergruppe zahlreiche psychometrische Besonderheiten diskutiert, die im Rahmen aktueller messmethodischer Debatten zur Kompetenzdiagnostik und zur Etablierung von Kompetenzmodellen interessante und relevante Fragen und Probleme betreffen. Daher verfolgt dieses Kapitel das Ziel, aktuelle methodische Herausforderungen zu beleuchten und einige verallgemeinerbare Schlussfolgerungen für das Large-Scale-Assessment abzuleiten.

Hierbei gliedert sich das Kapitel in zwei große Abschnitte: Abschnitt IV.2 behandelt Fragen der Dimensionalität der Lesekompetenz, in Abschnitt IV.3 wird das Problem des differenziellen Itemfunktionierens bei der Testung leistungsheterogener Gruppen diskutiert.

Konkret gehen wir in Abschnitt IV.2 auf methodisch relevante Aspekte der Konstruktdefinition der Lesekompetenz ein, welche sich insbesondere mit der Bestimmung der Konstruktdimensionalität und der Stabilität des Konstrukts beschäftigen. Die Frage, welche Subdimensionen der Lesekompetenz theoretisch plausibel unterschieden werden können und

inwieweit diese in bereits vorliegenden Studien empirische Unterstützung erfahren haben, birgt Implikationen für die Kompetenzdiagnostik in diesem Bereich ebenso wie für die Komplexität eines entsprechenden Kompetenzstufenmodells. Nach einer ausführlichen Gegenüberstellung verschiedener methodischer Zugänge zur Analyse der Konstruktdimensionalität im Abschnitt „Methode“ (IV.2.3) werden im Anschluss Ergebnisse für drei verschiedene analytische Zugänge vorgestellt.

Die Problematik, dass die Formulierung der Standards auf das Ende der Grundschulzeit, also die vierte Jahrgangsstufe,[1] abzielt, die künftigen Ländervergleiche ebenso wie die jährlichen Vergleichsarbeiten jedoch schon in der dritten Jahrgangsstufe durchgeführt werden, woraus sich gewisse methodische Schwierigkeiten ergeben, wird in Abschnitt IV.3 unter dem Stichwort des differenziellen Itemfunktionierens in leistungsheterogenen Gruppen thematisiert. Im Abschnitt „Befunde zur Verlinkung der Jahrgangsstufen drei und vier sowie zum differenziellen Itemfunktionieren“ präsentieren wir Befunde zur Verlinkung der Klassenstufen drei und vier und stellen verschiedene Ergebnisse zum differenziellen Itemfunktionieren auf Item- und Aufgabenebene vor. Eine Zusammenfassung und Diskussion dieser Ergebnisse erfolgt im nächsten Abschnitt.

Im Abschnitt IV.4 werden die beiden besprochenen Themen und die vorgestellten Befunde zusammengeführt. Es schließen sich eine Diskussion zu den angesprochenen methodischen Herausforderungen und Empfehlungen dazu an, wie ähnliche Probleme in verwandten Forschungskontexten gehandhabt werden könnten.

IV.2 Aspekte der Dimensionalität der Lesekompetenz im Primarbereich

IV.2.1 Die Konstruktdimensionalität und -stabilität der Lesekompetenz

Bremerich-Vos und Böhme (2009) erläutern, was im Rahmen der Operationalisierung der Bildungsstandards unter dem Begriff der Lesekompetenz zu verstehen ist und was mit dieser Bezeichnung in anderen großen Schulleistungsstudien sowie in der Deutschdidaktik verbunden wird. Eine interessante und relevante Frage, die bislang nicht thematisiert wurde, ist, ob sich auf theoretischer und/oder empirischer Ebene Teilaspekte der Lesekompetenz im Sinne von

[1] Mit Ausnahme der Länder Berlin, Brandenburg und zukünftig vermutlich auch Hamburg fällt das Ende der Grundschulzeit mit dem Ende der vierten Jahrgangsstufe zusammen.

trennbaren Subfacetten identifizieren lassen. Solche Subfacetten sind in der einschlägigen Literatur beispielsweise für verschiedene Textsorten oder „Leseintentionen", aber auch für verschiedene Verstehensaspekte diskutiert worden (für IGLU vgl. Bos, Valtin, Voss, Hornberg & Lankes, 2007; für PISA vgl. Schiefele, Artelt, Schneider & Stanat, 2004).

Ferner stellt sich die Frage, ob alle Aspekte der Lesekompetenz bereits im Primarbereich vollständig ausgeprägt und in gleichem Umfang relevant sind. Von Interesse ist hierbei, ob der Lernfortschritt eher quantitativer oder eher qualitativer Natur ist. Letzteres würde mit einer Veränderung der dimensionalen Struktur des Konstrukts über die Zeit hinweg einhergehen, Ersteres eher nicht. Auch verschiedene Unterrichtsschwerpunkte (Leseerwerb vs. literarische Leseerfahrung) und eine damit einhergehende differenzielle Förderung von Teilaspekten der Lesekompetenz (Bevorzugung literarischer Texte gegenüber kontinuierlichen und diskontinuierlichen Sachtexten in der Sekundarstufe I) könnten zu einer Veränderung der dimensionalen Struktur der Lesekompetenz führen.

Diese beiden Überlegungen sollen andeuten, dass Fragen der Konstruktdimensionalität wie auch der Konstruktstabilität über die Zeit hinweg wesentliche Bestandteile der Konstruktdefinition und der Entwicklung eines (theoretischen) Kompetenzmodells sind. Sie haben nicht nur Auswirkungen auf die Operationalisierung des Konstrukts und die Überprüfung der Kompetenzausprägung, sondern insbesondere auch auf die Förderung der Lesekompetenz im Unterricht (gezielt differenziell oder ganzheitlich) und die Dokumentation der Kompetenzstände in Kompetenzstufenmodellen.

Methodische Vorüberlegungen zur Dimensionalitätsprüfung

In der aktuellen empirischen Bildungsforschung besteht Konsens darüber, dass die genaue Charakterisierung des jeweils interessierenden Konstrukts unerlässlich ist. Ein wesentlicher Baustein einer solch umfassenden Validierung ist die Betrachtung der dimensionalen Struktur des operationalisierten Testkonstrukts (vgl. American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (NCME), 1999), welche sich aus dem Antwortverhalten der Testteilnehmer rekonstruieren lässt (vgl. Jang & Roussos, 2007; Tate, 2003). Üblicherweise wird die Dimensionalität eines Tests beziehungsweise die Dimensionalität des in ihm operationalisierten Konstrukts definiert als „the number of latent variables that account for the correlations among item responses in a particular data set" (Camilli, Wang & Fesq, 1995, S. 80). Tate definiert die Anzahl der Dimensionen eines Modells als „the minimum number of [...] latent abilities required to produce a monotone and locally independent model" (2003, S. 160) und fasst diese Eigenschaft als *strikte Dimensionalität*

(*Strict Dimensionality*) auf. Die Annahme der Eindimensionalität impliziert also, dass ein Test eine einzelne homogene Fähigkeit abbildet und die Items der Forderung nach lokaler stochastischer Unabhängigkeit genügen. Mehrdimensionalität (als Abweichung von strikter Eindimensionalität) bedeutet demnach, dass nicht ein einzelnes (dimensional) homogenes Konstrukt operationalisiert wurde. Als alternative Definition der Dimensionalität diskutiert Tate (2003) *essenzielle Dimensionalität* (*Essential Dimensionality*) im Sinne von Stout (1987, 2002). Hintergrund dieses Ansatzes ist das Bedürfnis, eine empirische Überprüfung des Grades der in einem Test realisierten (Ein-)Dimensionalität zu ermöglichen. Hierbei wird die Anzahl der Dimensionen so gewählt, dass die Annahme monotoner Item-Response-Funktionen und *essenzieller lokaler Unabhängigkeit* erfüllt ist. Stout nennt einen Test *essenziell eindimensional*, wenn der Mittelwert von Beträgen der Kovarianz von Itemresiduen eines Paares von Items (Residualkovarianzen) näherungsweise null ergibt (bzw. unabhängig von der Länge des Tests praktisch vernachlässigbar ist) und damit essenzielle lokale stochastische Unabhängigkeit im Mittel näherungsweise erfüllt ist (Stout, 1987, S. 597). Formal beschreibt man für Itemantworten X_i und eine eindimensionale Fähigkeit θ die Bedingung essenzieller Eindimensionalität als

$$\lim_{I \to \infty} \frac{\sum_{i<j} |Cov(X_i, X_j \,|\, \theta)|}{I \cdot (I-1)/2} \to 0$$

Dabei bezeichnen I die Itemanzahl sowie $\mathrm{Cov}(X_i, X_j | \theta)$ die nach der Fähigkeit θ bedingte Residualkovarianz. Die Formel besagt, dass der Mittelwert des Betrages der Residualkovarianzen aller Itempaare (i, j) für eine gegen unendlich gehende Itemzahl ($I \to \infty$) gegen den Wert null strebt. Damit kann dieses Kriterium als von der Itemanzahl unabhängige Effektgröße der Abweichung von Eindimensionalität angesehen werden. Diese oben definierte Größe wird im Fall von strikter Eindimensionalität den Wert null annehmen, da aufgrund der lokalen stochastischen Unabhängigkeit alle Residualkovarianzen gleich null sind. Formal ist die Definition nach Stout an einen unendlich großen Itempool gebunden, d. h. ein Test wird als essenziell eindimensional angesehen, wenn dieser für eine unendliche Itemanzahl im Mittel verschwindend geringe absolute Residualkovarianzen besitzt. Dass sich die erläuterten Annahmen der essenziellen Eindimensionalität auf eine unendliche Itemanzahl stützen, ist in der Literatur verschiedentlich kritisiert worden (vgl. Jannaronne, 1997). Allerdings gilt zu bedenken, dass ein Vorgehen, bei dem aus einem Universum unendlich vieler Items eine Itemstichprobe gezogen wird und für die resultierenden Itemparameter Messfehler (Standardfehler) geschätzt werden, beispielsweise den grundlegenden Annahmen der Generalisierbarkeitstheorie entspricht

(vgl. Brennan, 2001). Auch im Rahmen von Large-Scale-Assessments ist es plausibel, davon auszugehen, dass die Schülerinnen und Schüler eine Auswahl an Items bearbeiten, die exemplarisch für alle denkbaren Items der Operationalisierung eines Kompetenzkonstrukts stehen.

Vergleicht man beide Definitionen der Dimensionalität, so ist festzustellen, dass in beiden Perspektiven in gleicher Weise die Monotonie der *Item-Characteristic-Curves* (ICC) gefordert wird und der Unterschied ausschließlich im Hinblick auf die lokale stochastische Unabhängigkeit der Items besteht. Während bei strikter Dimensionalität das gesamte Pattern der Itemantworten, der Antwortvektor also, der Forderung nach lokaler stochastischer Unabhängigkeit entsprechen muss, bezieht sich diese Forderung bei essenzieller Dimensionalität lediglich auf Itempaare, die im Mittel „nur kleine" Abweichungen von lokaler stochastischer Unabhängigkeit zeigen dürfen. Die Forderung nach der lokalen Unabhängigkeit aller Items wird also der Forderung nach der „mittleren absoluten Abweichung von lokaler Unabhängigkeit von Itempaaren" gegenübergestellt. Für einzelne Itempaare sind bei essenzieller Eindimensionalität damit durchaus lokale stochastische Abhängigkeiten zugelassen.

Können im Fall der Mehrdimensionalität alle im Test enthaltenen Items so in distinkte Itemgruppen sortiert werden, dass jede Itemgruppe jeweils genau eine vorherrschende Dimension misst, so spricht man von einer Einfachstruktur (*Simple Structure*) der Daten (vgl. Stout et al., 1996), die die Interpretierbarkeit von Dimensionen deutlich vereinfacht. Unabhängig vom gewählten Ansatz ist die Anzahl ermittelter Dimensionen zwingend ganzzahlig, beispielsweise kann ein Modell nur ein- oder zweidimensional sein. Für in der Praxis vorliegende Datenkonstellationen wird man eine solche Aussage nicht immer in dieser Eindeutigkeit treffen können, sodass Alternativmodelle und ihre Sensitivität hinsichtlich substanzieller Aussagen überprüft und auch berichtet werden sollten. Auch werden theoretische Überlegungen bei der Modellwahl eine entscheidende Rolle spielen.

Stellen wir uns in einem Beispiel zwei korrelierte Dimensionen vor, deren Items eine Einfachstruktur aufweisen. Unabhängig von der Höhe der Korrelation wird man diese Daten im Ansatz der strikten Dimensionalität als zweidimensional auffassen. Im Rahmen des Konzepts essenzieller Dimensionalität sind zu .95 korrelierte Dimensionen jedoch „stärker essenziell eindimensional" als zu .70 korrelierte Dimensionen. Dies kann wie folgt heuristisch begründet werden:

Liegen auf beiden Dimensionen etwa gleich viele Items vor, so ist die „beste Schätzung" einer eindimensionalen Fähigkeit der Mittelwert aus beiden Dimensionen (wenn beide Dimensionen die gleiche Varianz besitzen). Werden nun im Ansatz essenzieller Dimensionalität Residualkovarianzen der jeweiligen Items – bedingt auf der gebildeten eindimensionalen

Schätzung – bestimmt, so fallen diese bei Items der gleichen Dimension positiv, bei Items verschiedener Dimensionen negativ aus. Die Residualkovarianzen zweier Items einer Dimension 1 nehmen nur dann den Wert null an, wenn bezüglich der wahren Dimension und nicht der mittleren Dimension bedingt wird. Im Fall einer hohen Korrelation von .95 ist jedoch die mittlere Dimension sehr ähnlich zur Dimension 1, sodass nur eine leicht positive Residualkovarianz entsteht. Für eine Korrelation von .70 wird die Residualkovarianz deutlich größer ausfallen. Mittelt man nun über alle Beträge dieser Residualkovarianzen, so entsteht ein größerer Betrag für den Test mit einer Korrelation von .70 zwischen den Dimensionen.

Die Stärke der Abweichung von essenzieller Eindimensionalität hängt aber im Allgemeinen nicht von der Anzahl der notwendigen Dimensionen im Sinne strikter Dimensionalität ab (Stout, 2002, S. 496). Ein zweidimensionaler Test kann eine deutlich stärkere Abweichung von Eindimensionalität als ein zehndimensionaler Test aufweisen. Gibt es in einem Lesetext Items zu neun Texten (*Testlets*), so kann dies zu einem zehndimensionalen Test führen, bei dem der erste Faktor als Generalfaktor der Lesekompetenz interpretiert werden kann. Die übrigen als unkorreliert angenommenen neun Dimensionen bilden textspezifische Faktoren ab. Zu verschiedenen Texten gehörige Items werden im Allgemeinen lokal stochastisch unabhängig sein und demzufolge Residualkovarianzen von null aufweisen (bzw. leicht negativ ausfallen). Die Beiträge der positiven Residualkovarianzen von Items eines Textes fallen dann bei der Definition der essenziellen Dimensionalität weniger stark ins Gewicht. Lässt man die Anzahl der Texte mit den zugehörigen Items gegen unendlich gehen, so entsteht in dieser Datenstruktur essenzielle Eindimensionalität, obwohl Items lokal stochastisch abhängig sind (für eine Kritik an dieser Konzeption vgl. Jannaronne, 1997). Dieser Ansatz berührt daher primär die Frage nach „praktisch relevanter Dimensionalität", sodass die von Walker, Azen und Schmitt (2006) gewählten eingängigen Bezeichnungen *Statistical* und *Substantive Dimensionality* anstelle strikter und essenzieller Dimensionalität gerechtfertigt erscheinen.

Stout argumentiert, dass „[…] the issue of unidimensionality is not addressed by showing the lack of fit of a particular unidimensional parametric family on IRT models to the given manifest distribution. Rather, the issue is whether any unidimensional IRT model exists that fits the manifest distribution" (2002, S. 488). Diese Aussage ist so zu verstehen, dass man in vielen Anwendungen vermutlich eher eine Überschätzung der Dimensionalität vorfindet. Die Nichtpassung eines (einparametrischen) eindimensionalen Rasch-Modells im Vergleich zu einem (einparametrischen) zweidimensionalen Rasch-Modell impliziert noch keine Mehrdimensionalität, wenn beispielsweise ein dreiparametrisches eindimensionales Modell eine gute Passung der Daten herstellen kann.

Aus methodischer Sicht können noch zwei weitere große Klassen von Ansätzen zur Überprüfung der Dimensionalität unterschieden werden: zum einen die Erfassung der vollen Dimensionalität (*Full Dimensionality*; vgl. Jang & Roussos, 2007), bei der festgelegt wird, wie viele Dimensionen ein gegebener Test operationalisieren soll und welche Items welcher Dimension zugeordnet werden, zum anderen der bloße Nachweis eines Mangels an Eindimensionalität.

Die Untersuchung der dimensionalen Struktur eines Testkonstrukts kann verschieden motiviert sein. Oftmals wird ein Forschungsinteresse an inhaltlicher Validierung und den beteiligten kognitiven Prozessen genannt (vgl. AERA, APA & NCME, 1999). Relevant kann auch die Frage sein, inwiefern die Annahme der Eindimensionalität, die der IRT durch das Postulat homogener Items zugrunde liegt, verletzt wird. Die Missachtung einer mehrdimensionalen Struktur in IRT-Analysen könnte beispielsweise zu einer Verschätzung der Itemschwierigkeiten oder einer Unterschätzung der Standardfehler der Itemschwierigkeiten führen. Bereits sehr früh in der Forschungstradition des Educational Measurement konnte aber gezeigt werden, dass IRT-Modelle gegenüber der Verletzung der Eindimensionalitätsannahme sehr robust sind (vgl. Reckase, 1979). Ferner konnten Reckase, Ackerman und Carlson (1988) an simulierten und realen Daten illustrieren, dass auch „mehrdimensionale" Items zu einem eindimensionalen Test zusammengefasst werden können. Dies gründet sich auf die Idee, dass Items praktisch niemals unverfälscht nur eine einzige Fähigkeit messen. Passt zum Beispiel ein dreidimensionales IRT-Modell gut auf vorliegende Daten und ist die Anzahl der Items auf den Dimensionen annähernd gleich, so argumentieren beispielsweise Kirisci, Hsu und Yu (2001), dass bei Vorliegen einer mehrdimensionalen Datenstruktur bei Korrelationen zwischen den Dimensionen größer als .40 die Anwendung eindimensionaler IRT-Modelle möglich ist. Mit dieser Argumentation ließe sich für zahlreiche Anwendungen die Wahl eindimensionaler Modelle rechtfertigen.

Ohnehin benötigt man zum Lösen eines Items in aller Regel ein Zusammenspiel mehrerer Kompetenzen, daher überprüfen Items meist mehr als eine einzige Fähigkeit. Dennoch können solche in diesem Sinne „mehrdimensionalen" Items dazu dienen, einen eindimensionalen Test zu konstruieren, der der Eindimensionalitätsannahme der IRT-Modelle zu entsprechen vermag. Voraussetzung hierfür ist allerdings, dass die Items jeweils die *gleiche* Zusammensetzung verschiedener Fähigkeiten operationalisieren, also die gleiche Komposition an Fähigkeiten messen. Steht nun die Einschätzung der Fähigkeit von Schülerinnen und Schülern im Vordergrund, können also auch Items, zu deren Lösung man verschiedene Kompetenzaspekte benötigt, verwendet werden, um einen eindimensionalen Test zu konstruieren. Nur dann, wenn die Struktur der kognitiven Kompetenz beziehungsweise ihrer Bestandteile von primärem

Interesse ist, kann auf die mehrdimensionale Modellierung der entsprechenden Facetten nicht verzichtet werden.

Empirische Befunde zur Konstruktdimensionalität der Lesekompetenz im Deutschen

Im Rahmen großer Schulleistungsuntersuchungen sind im deutschen Sprachraum bislang nur vereinzelt Versuche unternommen worden, die Dimensionalität der Lesekompetenz zu untersuchen. So liegen insbesondere Ergebnisse vor, die aus der PISA-Studie sowie der Internationalen Grundschul-Lese-Untersuchung (IGLU) gewonnen wurden. Im Rahmen der DESI-Studie wurde die Lesekompetenz nicht auf ihre Subdimensionen hin untersucht, es liegen lediglich Analysen zur übergreifenden Struktur aller erfassten sprachlichen Kompetenzen im Deutschen und Englischen vor (vgl. Jude et al., 2008). In der deutschdidaktischen Fachliteratur diskutiert Gailberger „Dimensionen und Teilanforderungen des Lesens" (2007, S. 149), benennt aber unter dem Begriff der Dimension Aspekte wie Vorwissen, Motivation, Interesse oder Emotion, die in der hier gewählten stärker psychologisch geprägten Perspektive eher als Kovariate der Lesekompetenz und nicht unbedingt als (konstruktimmanente) Dimensionen der Lesekompetenz verstanden werden. Unter Teilanforderungen des Lesens versteht Gailberger (2007) verschiedene Ebenen der kognitiven Informationsverarbeitung, wie etwa die Herstellung lokaler und globaler Kohäsion. Diese Differenzierung ähnelt den drei in PISA unterschiedenen Aspekten der Lesekompetenz, die dort als Dimensionen unterschieden werden.

Das bekannte und breit diskutierte Verständnis von Lesekompetenz der PISA-Studie des Jahres 2000, in der die Diagnostik der Lesekompetenz den Studienschwerpunkt darstellte, vertrat einen pragmatischen Ansatz, der sich bemühte, den Handlungsintentionen der Leserinnen und Leser gerecht zu werden (vgl. Artelt & Schlagmüller, 2004). Definiert wurde Leseverstehen in der PISA-Konzeption als die Fähigkeit „geschriebene Texte zu verstehen, zu nutzen und über sie zu reflektieren, um eigene Ziele zu erreichen [...]" (Baumert, Stanat & Demmrich, 2001, S. 23).

Im Rahmen von PISA 2000 wurde zunächst mit einem Modell gearbeitet, welches fünf trennbare Aspekte der Lesekompetenz postulierte. Diese wurden aufgrund der Befunde empirischer Analysen für die Berichtlegung auf die folgenden drei Dimensionen reduziert: „Informationen ermitteln", „textbezogenes Interpretieren" sowie „Reflektieren und Bewerten". Artelt und Schlagmüller (2004) zeigen, dass die messfehlerbereinigten Korrelationen zwischen diesen drei als Dimensionen gefassten Aspekten der Lesekompetenz zwischen $r = .88$ und $r = .94$ lagen. Die Autoren berichten außerdem Befunde aus weiterführenden Analysen zu PISA 2000 bezüglich der Differenzierbarkeit von Textgenres (literarische vs. nichtliterarische Texte) und Textformaten (kontinuierliche vs. nichtkontinuierliche geschriebene Texte). Es wird ein

(messfehlerbereinigter) Zusammenhang zwischen kontinuierlich geschriebenen und nichtkontinuierlich geschriebenen Lesetexten von r = .91 berichtet. Für die zusätzliche Untersuchung des Textgenres wählen Artelt und Schlagmüller (2004) eine dreidimensionale Struktur, die literarische Texte von einerseits kontinuierlichen und andererseits nichtkontinuierlichen Sachtexten abgrenzt. Dies erscheint sinnvoll, da auf diese Weise eine Konfundierung des Genres (literarischer Text vs. Sachtext) mit dem Format des Textes (kontinuierlich vs. nichtkontinuierlich) vermieden wird. Als Resultat ihrer Analysen präsentieren Artelt und Schlagmüller (2004) die in Tabelle IV.1 dargestellten (messfehlerbereinigten) Korrelationen (vgl. Artelt & Schlagmüller, 2004, S. 178).

Tabelle IV.1: Messfehlerbereinigte Korrelationen zwischen den Subskalen der Lesekompetenz bei literarischen, kontinuierlich und nichtkontinuierlich geschriebenen (Sach-)Texten (nach Artelt & Schlagmüller 2004, S. 178)

	Literarische Texte	Kontinuierliche Sachtexte
Kontinuierliche Sachtexte	.79	
Diskontinuierliche Sachtexte	.76	.90

Im Rahmen von IGLU 2006 werden Befunde aus mehrdimensionalen Rasch-Modellen vorgestellt, die Annahmen bezüglich zweier „Leseintentionen“ (literarische vs. informierende Texte) mit denen vier verschiedener Leseverstehensprozesse (Erkennen und Wiedergeben explizit angegebener Informationen; einfache Schlussfolgerungen ziehen; komplexe Schlussfolgerungen ziehen und begründen, Interpretieren des Gelesenen; Prüfen und Bewerten von Inhalt und Sprache) verknüpfen. Auf latenter Ebene bewegen sich die Korrelationen zwischen den vier Leseverstehensprozessen zwischen r = .92 und r = .98. Zum Zusammenhang der beiden „Leseintentionen“ findet sich die Aussage: „Die latenten Korrelationen [...] liegen um .9“ (Bos et al., 2007, S. 92).

Vergleichend zeigt sich, dass die in PISA 2000 gefundenen Zusammenhänge, die messfehlerbereinigt teils deutlich unter .80 liegen, eher auf trennbare Subdimensionen der Lesekompetenz hindeuten, als dies für die Befunde der IGLU-Studie der Fall ist. Mit Werten um oder über .90 weisen Letztere auf eine sehr starke Assoziiertheit der betrachteten Dimensionen hin.

Stabilität der Konstruktdimensionalität über den Erwerbsprozess

Eine relevante Frage, die sich an die oben dargestellten Überlegungen zur Konstruktdimensionalität anschließt, ist die nach der Stabilität eines Konstrukts über die Zeit und somit über Entwicklungsstände etwa in der Grundschule und Sekundarstufe I hinweg.

Einerseits ist verständlich, dass die Tatsache, dass Schülerinnen und Schüler hinzulernen, somit ihr Wissen und ihre Kompetenz stetig erweitern, dazu führt, dass sie beispielsweise in der Sekundarstufe nicht nur (quantitativ) mehr, sondern auch (qualitativ) anderes wissen und können als während ihrer Grundschulzeit. Dies ist erklärtes Ziel der schulischen Bildung. So kann man fragen, warum und inwiefern sich die Lesekompetenz im Hinblick auf recht klar umrissene Konstruktaspekte, wie gezielt Informationen zu identifizieren, Gelesenes mit dem eigenen Vorwissen zu verknüpfen, Inferenzen zu bilden, zu interpretieren und zu bewerten, qualitativ verändert oder ob nicht doch lediglich das Ausmaß des Könnens zunimmt.

Denkbar ist, dass unter denselben begrifflichen Bezeichnungen wie „Interpretieren" oder „Bewerten" in Grundschule und Sekundarstufe ganz verschiedene kognitive Prozesse gefasst werden, da qualitative Veränderungen des Konstrukts stattfinden. Eine alternative Perspektive könnte davon ausgehen, dass die ablaufenden kognitiven Prozesse im Wesentlichen identisch sind, aber die Gegenstände, an denen sie sich vollziehen, mit der Zeit immer komplexer werden. Dies würde eher eine quantitative Veränderung bedeuten.

Möglich wäre ferner eine Veränderung der dimensionalen Struktur des Konstrukts aufgrund verschiedener Schwerpunkte im Unterricht. Während der Fokus im Primarbereich auf dem Lesenlernen liegt, verschiebt sich dieser Schwerpunkt in der Sekundarstufe I hin zum Literaturunterricht. Die Förderung der Lesekompetenz erfolgt somit eher differenziell. Verglichen mit kontinuierlichen und diskontinuierlichen Sachtexten erfahren literarische Texte im Unterricht eine deutlich intensivere Zuwendung (vgl. Spinner, 2004)[2].

Ob und, falls ja, welche Unterschiede in der dimensionalen Struktur des Konstrukts der Lesekompetenz über die Schulzeit hinweg zu erwarten sind, ist aus unserer Sicht keine triviale Frage. Eine vertiefte Klärung ist uns an dieser Stelle jedoch nicht möglich, da die vorliegenden Daten keine längsschnittliche Betrachtung gestatten. Deutlich wird aber, dass mit der Diskussion um die Stabilität der dimensionalen Struktur der Lesekompetenz auch die Frage einhergeht,

[2] In fachdidaktischen Kreisen wird jedoch berichtet, dass sich diesbezüglich Änderungen im Unterrichtsgeschehen abzeichnen und – im Zuge von PISA und der Ermittlung spezifischer Leistungsdefizite von Schülern gegenüber Schülerinnen – Sachtexten in der Sekundarstufe I vermehrt Aufmerksamkeit geschenkt wird.

inwieweit die Lesekompetenzleistungen der Schülerinnen und Schüler in Grundschule und Sekundarstufe I miteinander vergleichbar sind.

IV.2.2 Fragestellung

Vor dem Hintergrund der hier thematisierten Problembereiche stellt sich für uns die folgende Frage: Lassen sich im Primarbereich verschiedene Subdimensionen der Lesekompetenz, die sich auf für die Sekundarstufe bereits etablierte Differenzierungen (vgl. Artelt & Schlagmüller, 2004) beziehen, unter Rückgriff auf verschiedene methodische Zugänge nachweisen?

IV.2.3 Methode

Beschreibung der Personen- und Itemstichprobe

Die nachfolgenden Analysen beziehen sich auf Daten, die im Rahmen der Normierungsstudie des Jahres 2007 gewonnen wurden. Im Rahmen dieser Studie bearbeiteten 3 600 Schülerinnen und Schüler (davon 1 706 aus Jahrgangsstufe drei und 1 894 aus Jahrgangsstufe vier) aus 112 dritten und 112 vierten Klassen Aufgaben zur Überprüfung ihrer Lesekompetenz. Hierfür wurden 19 Testaufgaben mit insgesamt 113 Items eingesetzt. Bemüht man sich um eine Zuordnung der Aufgaben zu den oben diskutierten möglichen Subfacetten der Lesekompetenz, so messen acht Aufgaben das Leseverstehen im Hinblick auf kontinuierliche, literarische Texte. Sechs Aufgaben erfassen die Lesekompetenz bei kontinuierlichen Sachtexten und fünf Aufgaben bei diskontinuierlichen Sachtexten.

Methodische Zugänge zur Analyse der Konstruktdimensionalität

In der einschlägigen Literatur werden verschiedene Ansätze zur Bestimmung der dimensionalen Struktur des Antwortverhaltens zu operationalisierten Testkonstrukten vorgestellt (vgl. Ackerman, Gierl & Walker, 2003; vgl. Jang & Roussos, 2007; vgl. Tate, 2003).

Wir möchten an dieser Stelle eine kurze Zusammenschau einiger methodischer, in der Literatur diskutierten Zugänge zur Analyse der Konstruktdimensionalität vorstellen. Für einen systematischen Überblick lassen sich für dichotome Itemantworten gemäß Tate (2003) zunächst parametrische von nichtparametrischen Ansätzen abgrenzen. Während in parametrischen Ansätzen die Parameter eines Populationsmodells geschätzt werden, liegt der Analyse in nichtparametrischen Zugängen kein (endlichdimensional) parametrisiertes, zu schätzendes

Populationsmodell zugrunde. Beispielsweise muss im Modell nicht die Höhe der Korrelation zwischen zwei distinkten Itemgruppen, die verschiedene Dimensionen messen, als Parameter vorliegen. Die verschiedenen Ansätze sind in Abbildung IV.1 zusammengefasst. Viele der dargestellten Zugänge können sowohl in konfirmatorischer, also Hypothesen prüfender, als auch in explorativer, also Hypothesen generierender Absicht verwendet werden.

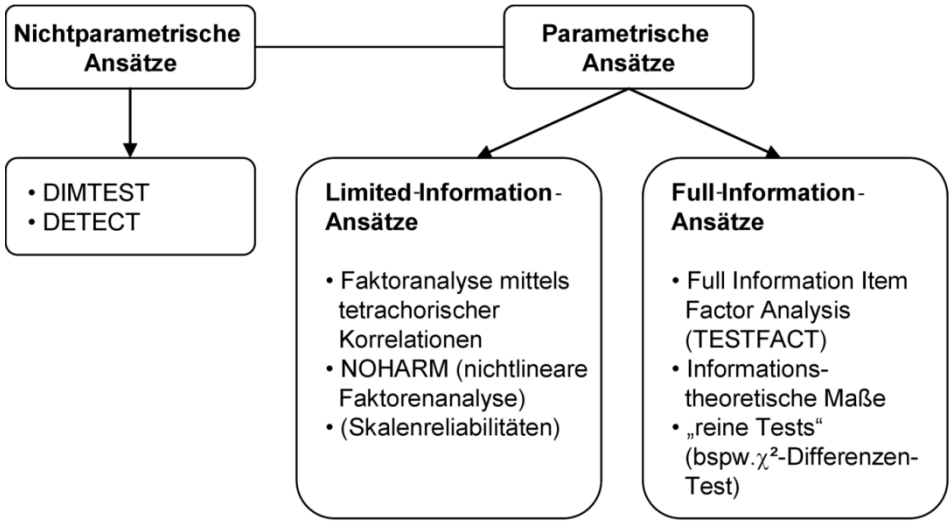

Abbildung IV.1: Auswahl parametrischer und nichtparametrischer Zugänge zur Prüfung der Konstruktdimensionalität

Neben Tate (2003) liefern Ackerman et al. (2003) einen sehr lesenswerten Überblicksartikel, der die Mehrdimensionalität eines Tests auch im Hinblick auf praxisrelevante Implikationen diskutiert. In ihrem Beitrag konzentrieren sich die Autoren schwerpunktmäßig auf mehrdimensionale IRT-Modelle. In der von uns gewählten Klassifizierung methodischer Ansätze zur Analyse der Konstruktdimensionalität behandeln wir mehrdimensionale IRT-Modelle (*MIRT-Modelle*) nicht als geschlossenen, eigenständigen Zugang. Vielmehr sind MIRT-Modelle ebenso die Grundlage für die Anwendung von Modelltests beispielsweise mittels informationstheoretischer Maße, wie wir sie unter dem Stichwort der *Full-Information-Ansätze* behandeln, wie auch für die Anwendung von NOHARM, welches als *Limited-Information-Ansatz* die Itemparameter mehrdimensionaler IRT-Modelle schätzen kann. Die bloße Einschätzung der Höhe der in einem mehrdimensionalen IRT-Modell erhaltenen Korrelationen auf latenter Ebene sollte jedoch nicht als alleiniger Indikator zur Überprüfung der dimensionalen Struktur im Sinne eines Modelltests dienen.

Ferner stellen faktorenanalytische Verfahren für dichotome Daten spezielle multidimensionale IRT-Modelle dar. Für einen integrierenden Überblick zu den Beziehungen dieser Modellklassen sei auf Kamata und Bauer (2008) verwiesen (vertieft vgl. auch Skrondal & Rabe-Hesketh, 2004).

Nichtparametrische Ansätze

Im Rahmen nichtparametrischer Ansätze stehen die Verfahren DETECT (vgl. Zhang & Stout, 1999; vgl. Zhang, 2007) und DIMTEST (vgl. Stout, 1987; vgl. Stout, Froelich & Gao, 2001) zur Verfügung. Beide Verfahren beruhen auf dem Ansatz essenzieller Eindimensionalität und verwenden auf bedingten Residualkovarianzen basierende Maße als Teststatistiken. Die DETECT-Statistik kann dabei als Schätzung eines Populationskennwertes im Hinblick auf die Facetten Personen und Items verstanden werden (vgl. Roussos & Ozbek, 2006). Wählt man einen konfirmatorischen Zugang mittels DETECT, so müssen zunächst eindeutige Klassifikationen der Items entsprechend den postulierten (Sub-)Kompetenzen, in unserem Fall entsprechend den verschiedenen theoretisch fundierten Subdimensionen der Lesekompetenz, vorgenommen werden. Diese A-priori-Zuordnung der Items zu den postulierten Dimensionen dient als Grundlage, um die Abweichung dieser Itemgruppierung von der Eindimensionalität zu untersuchen. Quantifiziert wird diese Abweichung über verschiedene Prüfgrößen. Diese sind der DETECT-Index, der ASSI-Wert (*Approximative Simple Structure Index*), welcher bei Jang und Roussos (2007) als IDN-Index bezeichnet wird, sowie der RATIO-Index.

In der Literatur vorgenommene Interpretationen der Ausprägungen dieser Indizes können Tabelle IV.2 entnommen werden. Deutlich wird, dass die als kritisch diskutierten Richtwerte für den DETECT-Index in jüngerer Zeit etwas liberaler ausfallen als in früheren Arbeiten (vgl. Ackermann et al., 2003; vgl. Jang & Roussos, 2007; vgl. Kim, 1994 zitiert nach Ackermann et al., 2003). Verwendet man DETECT für explorative Analysen, so wird keine A-priori-Zuweisung der Items zu möglichen Subdimensionen vorgenommen, allerdings wird die Anzahl zu prüfender Dimensionen vorgegeben. Üblicherweise wählt man hierbei nicht nur eine Vorgabe, sondern variiert die Anzahl der Dimensionen innerhalb eines theoretisch plausiblen Bereichs. Das DIMTEST-Verfahren kann sowohl für konfirmatorische Zwecke als auch für explorative Zugänge verwendet werden. Mitunter wird in der Literatur zunächst DETECT zur Hypothesengenerierung und anschließend DIMTEST zur Überprüfung dieser Hypothese verwendet (vgl. Jang & Roussos, 2007).

Tabelle IV.2: In der Literatur berichtete Interpretationen der in DETECT verwendeten Prüfgrößen

DETECT-Index		
Wertebereich nach Jang & Roussos (2007)	Wertebereich nach Kim (1994), zitiert nach Ackerman et al., (2003)	Interpretation
> 1	> 1	Starke Mehrdimensionalität
.4 – 1	.5 – 1	Moderate bis starke Mehrdimensionalität
.2 – .4	.1 – .5	Schwache Mehrdimensionalität
< .2	< .1	Essenzielle Eindimensionalität
ASSI (Approximative Simple Structure Index) (IDN)		
1	Maximale Ausprägung der dimensionalen *simple structure*	Eigene Darstellung unter Bezugnahme auf Zhang (2007)
> .25	Abweichung von Eindimensionalität	
< .25	Essenzielle Eindimensionlität	
RATIO-Index		
1	Maximale Ausprägung der dimensionalen *simple structure*	Eigene Darstellung unter Bezugnahme auf Zhang (2007)
> .36	Abweichung von Eindimensionalität	
< .36	Essenzielle Eindimensionlität	

Parametrische Ansätze

Bei den parametrischen Ansätzen unterscheiden wir *Limited-Information-Ansätze* von *Full-Information-Ansätzen* und wollen uns zunächst den Limited-Information-Ansätzen zuwenden, die auch als „klassische Ansätze“ bezeichnet werden könnten. Diese basieren auf paarweisen Itembeobachtungen, sodass für die Analysen zweidimensionale Kontingenztabellen für alle Itempaare ausreichen. Einen Zugang zur Bestimmung der Konstruktdimensionalität im Rahmen dieser Klasse von Methoden bilden lineare Faktorenanalysen, die auf der Grundlage tetrachorischer Korrelationen berechnet werden. Ein sehr einfacher Zugang ist die grafische Prüfung des Eigenwertverlaufs (*Scree-Plot*), anhand dessen oftmals ein klar dominanter erster Faktor identifiziert werden kann, der einen Großteil der Itemvarianz aufklärt. Eine Erweiterung dieses Zugangs kann in der Gegenüberstellung des Eigenwertverlaufs einer theoretisch fundierten Zuordnung zu dem einer Zufallszuordnung bestehen (Parallelplot, zitiert in Hattie, 1985). Wählt

man die Anzahl der Dimensionen stattdessen aufgrund des Prozentanteils aufgeklärter Varianz oder aufgrund des Quotienten zweier aufeinanderfolgender Eigenwerte (vgl. Hattie, 1985), so folgt man eher der Idee essenzieller Eindimensionalität. Man wird sich für ein eindimensionales Modell entscheiden, wenn der Varianzanteil des ersten Faktors und das Varianzverhältnis aus erstem und zweitem Eigenwert möglichst groß ausfallen.

Für dichotome Daten wurde in der Literatur eine modifizierte Parallelanalyse vorgeschlagen, die als zu vergleichende Eigenwerte Daten aus einem eindimensionalen IRT-Modell simuliert (vgl. Finch & Monahan, 2008). Ähnlich wie bei der Hauptkomponentenanalyse kann so auch für dichotome Daten ein Parallelplot eingesetzt werden, der am Datensatz ermittelte Eigenwerte mit simulierten Eigenwerten unter Annahme unkorrelierter Variablen vergleicht. Ist dabei der zweite beobachtete Eigenwert kleiner als der zweite simulierte Eigenwert, so ist von Eindimensionalität auszugehen. Alternativ dazu kann man auch das Verhältnis aus zweitem beobachteten und simulierten Eigenwert einsetzen, das bei Eindimensionalität nicht deutlich größer als eins ausfallen soll.

Ein Problem, das bei Faktorenanalysen tetrachorischer Korrelationen auftreten kann, ist eine Abweichung von der Normalverteilungsannahme der latenten Fähigkeit, deren Dimensionalität geprüft werden soll, die jedoch in vielen IRT-Modellen als gegeben vorausgesetzt wird. Ebenso können bei moderaten Stichprobenumfängen nicht positiv definite Kovarianzmatrizen resultieren, die zu Schätzinstabilitäten führen (vgl. Tran & Formann, 2009). Ein Nachteil hier aufgeführter Limited-Information-Methoden besteht darin, dass diese oft keine in Multi-Matrix-Designs vorliegenden Daten erlauben und damit für Large-Scale-Assessments nur eingeschränkt Anwendung finden können.

Weniger geeignet, in der Literatur jedoch als Möglichkeit zur Feststellung der Konstruktdimensionalität thematisiert, sind Maße der Skalenreliabilität (vgl. Blais & Laurier, 1995). Die Höhe der berechneten Reliabilitätsindizes wird geringer ausfallen, wenn es sich nicht um ein homogenes, eindimensionales Konstrukt handelt. Problematisch scheint in diesem Zusammenhang, dass die Höhe der Skalenreliabilität auch wesentlich durch die Anzahl der einbezogenen Items bestimmt wird. Ebenso können geringe Trennschärfen einzelner Items einerseits Ausdruck mangelhafter Itemkonstruktion sein und andererseits als Indiz dafür gelten, dass sie nicht die im Test vordergründige Eigenschaft erfassen. Den Problemgehalt der Berücksichtigung von Maßen der Skalenreliabilität veranschaulichen beispielsweise Borg und Staufenbiel (2007). Sie geben zu bedenken, dass eine „mechanische“ Itemanalyse, die ausschließlich auf die Maximierung von Cronbachs α abzielt, zu völlig falschen Entscheidungen bei der Itemselektion und auch bei der Interpretation der verbleibenden Skala führen kann. Diesen Umstand illustrieren sie sehr pointiert am Beispiel eines Intelligenztests:

> Intelligenz ist ein mehrdimensionales Phänomen und die aus einer blinden Maximierung von α durch Elimination inhomogener Items letztlich erzeugte Skala somit entweder nur ein fauler Kompromiss, der Äpfel und Birnen [...] zusammenwirft, oder ein reines Birnenkompott, weil alle Äpfel eliminiert wurden. (Borg & Staufenbiel, 2007, S. 395)

In der Klasse der parametrischen Full-Information-Ansätze kann zunächst die explorative oder konfirmatorische Verwendung der auf vollständigen Item Responses basierenden *Full-Information-Item-Factor-Analyses* (TESTFACT; vgl. Bock, Gibbons & Muraki, 1988; für einen Überblick siehe Wirth & Edwards, 2007) genannt werden. Ferner bilden mehrdimensionale Modellierungen der probabilistischen Testtheorie (IRT) einen weiteren möglichen Zugang innerhalb dieser Klasse. Als Prüfgrößen der (strikten) Dimensionalität dienen hier in erster Linie informationstheoretische Indizes, mit deren Hilfe Modellgütevergleiche durchgeführt werden können (vgl. Burnham & Anderson, 2002). Hierbei wird die Güte der Anpassung der empirischen Daten an ein theoretisch postuliertes Modell bestimmt, für das die Likelihood-Funktion berechnet wird. In die Bestimmung der Anpassung fließen die maximale Likelihood sowie die Anzahl der beinhalteten Modellparameter als Maß der Sparsamkeit ein. Die gängigen Informationskriterien sind der AIC (*Akaikes Information Criterion*), der BIC (*Bayes Information Criterion*; auch SIC, *Schwarz Information Criterion*) sowie der CAIC (*Consistent AIC*). Die Informationskriterien können für ein einzelnes Modell nicht in ihrer absoluten Höhe interpretiert werden, sondern nur im Vergleich verschiedener Modelle, die gegeneinander getestet werden. Hierbei stellt das Modell mit den geringeren Werten der Informationskriterien das besser fittende Modell dar. Allerdings haben diese Statistiken den Nachteil der Abhängigkeit vom Stichprobenumfang der Personen (vgl. Marsh, Balla & McDonald, 1988; vgl. McDonald & Mok, 1995) und dem Stichprobenumfang der Items, sodass häufig bei großen Stichproben komplexere (höherdimensionale) Modelle präferiert werden. Liegen Items in einem Multi-Matrix-Sampling vor, dann stellt sich jedoch die Frage, welche Stichprobengröße für die Berechnung des BIC und des CAIC herangezogen werden soll (vgl. McCoach & Black, 2008).

Weitere Varianten zur Prüfung der Modellgüte sind „reine" Tests der Modellpassung, beispielsweise der Likelihood-Quotienten- beziehungsweise der $\chi 2$-Differenzen-Test.

IV.2.4 Befunde der Dimensionalitätsanalysen

Entsprechend der in Abschnitt „Methodische Vorüberlegungen zur Dimensionalitätsprüfung" erläuterten Unterscheidung bemühen wir uns im Folgenden um die methodische Erfassung der vollen Dimensionalität, da wir vorab theoriegeleitet bestimmen, wie viele Dimensionen

unterschieden werden und darüber hinaus eine eindeutige Zuordnung der Items zu den Dimensionen vornehmen.

Hierfür möchten wir Ergebnisse sowohl parametrischer als auch nichtparametrischer Analysen vorstellen, wobei wir uns bei allen Ergebnissen immer auf die Analyse dichotomer Daten beziehen. Zunächst berichten wir Befunde, die sich aus einer Implementierung des DETECT-Verfahrens in der Software R (vgl. R Development Core Team, 2008) ergaben. Anschließend werden wir auf die faktorielle Untersuchung der Dimensionalität auf Aufgabenebene und die Ergebnisse aus mehrdimensionalen IRT-Analysen eingehen.

Befunde nichtparametrischer Ansätze (DETECT)

Wählt man für DETECT einen *konfirmatorischen Ansatz*, können die folgenden drei theoretisch plausiblen und in der einschlägigen Literatur (vgl. Artelt & Schlagmüller, 2004) diskutierten Dimensionen des Leseverstehens unterschieden werden: 1) kontinuierliche literarische Texte, 2) kontinuierliche Sachtexte sowie 3) diskontinuierliche Sachtexte. Für diese dreidimensionale Differenzierung der Lesekompetenz ergibt sich ein DETECT-Wert von .24, der entsprechend den oben dargestellten Interpretationsvorschlägen von Jang und Roussos (2007) eine schwache Mehrdimensionalität anzeigt. Für die Indizes ASSI und RATIO ergeben sich jeweils Werte von .11, die in ihrer Höhe beide als Anzeichen einer essenziellen Eindimensionalität gewertet werden können (vgl. Zhang, 2007). Diese Befunde lassen keine zweifelsfreie Entscheidung zugunsten einer ein- oder einer mehrdimensionalen Lösung zu. Es wird jedoch deutlich, dass die jüngeren, eher liberaleren Grenzwerte für den DETECT-Index möglicherweise weniger gut mit der Interpretation der anderen beiden Indizes (ASSI und RATIO) korrespondieren, als es die strengeren, heute aber weniger gebräuchlichen Grenzen (vgl. Kim, 1994) tun.

Wählt man ergänzend auch ein *explorativ*es *Vorgehen* in DETECT, bei dem zwar keine A-priori-Zuordnung der Items zu möglichen Subdimensionen erfolgt, aber die Anzahl der zu prüfenden Cluster vorgegeben wird, die in unserem Fall auf drei Cluster festgelegt wurde, so zeigen sich etwas stärkere Hinweise auf eine mögliche Mehrdimensionalität, wobei diese in erster Linie auf die Ausprägung des DETECT-Wertes zurückgeführt werden können. Dieser indiziert mit einem Betrag von .74 moderate bis starke Mehrdimensionalität. Demgegenüber liegen sowohl der ASSI als auch der RATIO-Index unterhalb der kritischen Grenzwerte, die eine Abweichung von essenzieller Eindimensionalität kennzeichnen. Die für das konfirmatorische und explorative Vorgehen erhaltenen Befunde der Dimensionalitätsprüfung der Lesekompetenz mittels DETECT sind in Tabelle IV.3 nochmals zusammengefasst.

Es muss betont werden, dass die explorativ erhaltene Klassifikation der Items in drei Cluster, also drei Dimensionen, nicht der konfirmatorischen A-priori-Klassifikation entspricht. Falls die explorativen DETECT-Befunde als Indiz für Mehrdimensionalität gedeutet werden sollen, müsste für eine inhaltlich sinnvolle Interpretation dieser Dimensionen also zunächst eine fachdidaktisch und kognitionspsychologisch informierte Analyse der erhaltenen Itemklassifikation vorgenommen werden, da die explorativ entstandene Klassifikation nicht mit der theoretisch postulierten identisch ist.

Zusammenfassend kann mithilfe des hier eingesetzten nichtparametrischen Verfahrens DETECT keine eindeutig mehrdimensionale Struktur identifiziert werden, die einer Replikation der für den Sekundarstufenbereich berichteten dimensionalen Struktur der Lesekompetenz entsprechen würde. Dies muss jedoch den bislang in der Literatur berichteten Befunden nicht widersprechen, vielmehr liefert diese Methode weitere Anhaltspunkte, die mit Ergebnissen aus anderen Studien in Zusammenhang gebracht werden müssen.

Tabelle IV.3: Befunde einer konfirmatorischen und explorativen Dimensionalitätsprüfung mittels DETECT

DETEC-Index	Konfirmatorisches Modell (*A-priori*-Klassifikation der Items in 3 Teildimensionen)	Exploratives Modell (keine *A-priori*-Klassifikation der Items; 3 Cluster)
DETECT	.24	.73
ASSI (IDN)	.11	.15
RATIO	.11	.26

Untersuchung der Dimensionalität auf Aufgabenebene mittels parametrischer Limited-Information-Ansätze

Leseaufgaben zeichnen sich dadurch aus, dass jeweils mehrere Items zu einem gemeinsamen Stimulus (Lesetext; Testlet) gehören. Aus dieser Verwandtschaft der Items, die sich jeweils auf denselben Text beziehen, kann eine Verletzung der lokalen stochastischen Unabhängigkeit resultieren. Um zu ermitteln, wie stark die Abhängigkeiten der Items eines Textes sind, können so genannte Testletmodelle eingesetzt werden (vgl. Robitzsch, 2009). Zusätzlich besteht das Problem, dass auch Texte verschiedener Textsorten konstruktrelevante Varianz binden und somit eine Verzerrung der ermittelten dimensionalen Struktur erzeugen können. Für eine Bestimmung der Dimensionalität des Konstrukts nach Kontrolle für das Stimulusmaterial setzen

wir daher keine Testletmodelle ein, sondern auf tetrachorischen Korrelationen basierende Faktorenanalysen für jede der 16 Aufgaben der Normierungsstudie des Jahres 2007.

Jede Aufgabe greift auf genau einen Stimulus zurück. Bei 10 von 16 Aufgaben ergibt sich in einem Parallelplot eine Einfaktorstruktur. Das Verhältnis des zweiten beobachteten Eigenwertes und des zweiten Eigenwertes aus einer Zufallskorrelationsmatrix (für einen Parallelplot) ist nur für zwei Lesetexte substanziell größer als 1.1, sodass die Items dieser beiden Aufgaben nicht als eindimensional aufgefasst werden können. Der durch den ersten Faktor aufgeklärte Varianzanteil beträgt für die 16 Aufgaben im Mittel 48,2 % mit einer Standardabweichung von 14,1 (Minimum: 32,4 %, Maximum: 83,4 %). Ebenso indiziert das Varianzverhältnis aus erstem und zweitem Eigenwert mit einem Mittelwert von 3.12 und einer Standardabweichung von 1.46 (Minimum: 1.24, Maximum: 7.63) (essenzielle) Eindimensionalität. Wir finden also einen dominanten ersten Faktor für die 16 eingesetzten Leseaufgaben. Aus der Perspektive eines Testletmodells wird mit diesem Vorgehen jedoch nur gezeigt, dass die Interaktionen von Schüler und Testlet jeweils eindimensionale Fähigkeiten darstellen.

Befunde parametrischer Ansätze unter Rückgriff auf Full-Information-Maximum-Likelihood-Verfahren

Die in Abbildung IV.2 dargestellten Zusammenhänge auf latenter, also messfehlerbereinigter Ebene weisen in eine ähnliche Richtung wie Befunde, die im Rahmen vertiefender Analysen der PISA-Studie 2000 von Artelt und Schlagmüller (2004) berichtet wurden (vgl. Abschnitt IV.2.1).

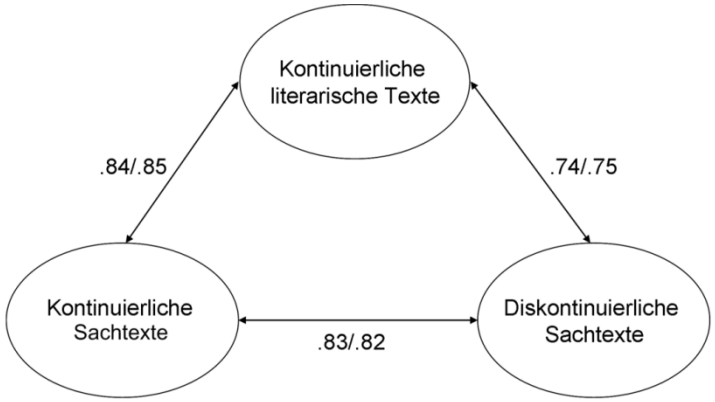

Abbildung IV.2: Korrelationen möglicher Subdimensionen der Lesekompetenz bei literarischen Texten sowie kontinuierlichen und diskontinuierlichen Sachtexten

Anmerkung: Die Werte bezeichnen messfehlerkorrigierte Korrelationen auf latenter Ebene in der Form: Korrelation aus dreidimensionaler Skalierung in ConQuest/Korrelation aus paarweiser zweidimensionaler Skalierung in Mplus.

Die dargestellten Zusammenhänge ergeben sich für eine gemeinsame Analyse der Daten der dritten und vierten Jahrgangsstufe[3]. Es zeigt sich deutlich, dass die Wahl verschiedener Skalierungsmethoden (einparametrische, gemeinsame dreidimensionale Skalierung in ConQuest [vgl. Wu, Adams, Wilson & Haldane, 2007] vs. zweiparametrische, paarweise zweidimensionale Skalierung in Mplus [vgl. Muthén & Muthén, 2006]) nicht zu praktisch relevanten Unterschieden in den Befunden führt. Die Höhe der Zusammenhänge weist eine ähnlich starke Beziehung für kontinuierliche literarische Texte mit kontinuierlichen Sachtexten und für kontinuierliche mit diskontinuierlichen Sachtexten aus. Der Zusammenhang zwischen kontinuierlichen literarischen Texten und diskontinuierlichen Sachtexten ist etwas niedriger.

Die Größenordnung der hier gefundenen Zusammenhänge ist ähnlich der von Artelt und Schlagmüller (2004) berichteten, wenn auch die Reihenfolge der Beziehungen von der stärksten zur schwächsten nicht identisch ist. Zu konzedieren ist, dass die Korrelationen allesamt sehr hoch ausfallen und man annehmen kann, dass die gemeinsame Varianz der drei Facetten auf einen generellen Lesefaktor zurückzuführen ist.

Um eine Aussage über die dimensionale Struktur treffen zu können, liefert die Höhe der Korrelationen allein allerdings keine hinreichende Information. Daher werden in Tabelle IV.4

[3] Eine getrennte Analyse für beide Jahrgangsstufen sowie die Prüfung der Invarianz der identifizierten Strukturen über die Jahrgänge hinweg sind an dieser Stelle nicht möglich, da die nach Jahrgängen getrennte, pro Textsorte verfügbare Anzahl an Aufgaben und Items unzureichend ist.

auch die Resultate eines Modellgütevergleichs der ein- sowie dreidimensionalen Skalierungen zusammengefasst. Gemessen an den Informationskriterien zeigt sich eine bessere Passung des mehrdimensionalen Modells.

Tabelle IV.4: Modellgütevergleich ein- und dreidimensionaler Skalierungen in ConQuest

	Parameteranzahl	Devianz	AIC	BIC
Eindimensional Modell	113	61346	61572	61748
Dreidimensionales Modell	119	61067	61305	61490

Anmerkung: AIC = Akaikes Information Criterion, BIC = Bayes Information Criterion

IV.2.5 Zusammenfassung und Diskussion

Abschließend können wir an dieser Stelle festhalten, dass die präsentierten Ergebnisse zur Konstruktdimensionalität der Lesekompetenz im Primarbereich dahingehend interpretiert werden können, dass zwar keine klar eindimensionale Struktur identifiziert wurde, aber auch keine eindeutig mehrdimensionale Struktur vorliegt, die eine Ableitung jeweils eigenständiger Kompetenzstufenmodelle für die verschiedenen Subfacetten gestatten würde. Mittels der konzeptuell verschiedenen Zugänge der strikten und essenziellen Dimensionalität sind die Befunde auf den ersten Blick verschieden. Bezüglich der strikten Dimensionalität sprechen Modellgütetests mehrdimensionaler IRT-Modelle für das dreidimensionale Modell. Allerdings ergeben sich auf latenter Ebene moderate bis hohe Korrelationen zwischen verschiedenen Textsorten, die in Kombination mit den auf essenzieller Eindimensionalität beruhenden Verfahren praktisch dafür sprechen, bei der Modellierung der Lesekompetenz in sinnvoller Näherung mit eindimensionalen Modellen zu arbeiten.

Die hier vorgestellten, nach der Textsorte differenzierten dreidimensionalen Modelle sind hinsichtlich der spezifischen Varianz in den Dimensionen zu erweitern. In Verbindung der Tradition der 3-Ebenen-Modelle (vgl. Raudenbush & Bryk, 2002) und hierarchischer konfirmatorischer Faktorenanalysen sind IRT-Modelle der Lesekompetenz denkbar, in denen die Gesamtvarianz in einen Generalfaktor der Lesekompetenz (Ebene 3), in die drei textsortenspezifischen Faktoren (auf Ebene 2) und testletspezifische Faktoren (auf Ebene 1)

zerlegbar wäre. Schülerleistungen in Testlets wären demnach geschachtelt in spezifischen Schülerleistungen in Textsorten, die wiederum in die allgemeine Schülerleistung im Leseverstehen eingebettet sind. Die Erfassung der Varianzanteile auf den verschiedenen Ebenen könnte dann ein komplettierendes Bild der Erfassung der Lesekompetenz liefern.

IV.3 Differenzielles Itemfunktionieren in der dritten und vierten Jahrgangsstufe

IV.3.1 Bildungspolitische Ausgangssituation

Die Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Kultusministerkonferenz, KMK) hat im Sommer des Jahres 2006 eine Gesamtstrategie zum Bildungsmonitoring verabschiedet, in der auch die zentrale Überprüfung des Erreichens der Bildungsstandards thematisiert wird (vgl. KMK, 2006). Im Rahmen dieser Gesamtstrategie wird deutlich, dass die Bildungsstandards im Fach Deutsch für den Primarbereich (vgl. KMK, 2005) als Leistungserwartungen formuliert wurden, die sich auf das Ende der vierten Jahrgangsstufe beziehen. Die künftige Überprüfung des Erreichens dieser Leistungserwartungen soll aber bereits ein Jahr früher stattfinden. So heißt es in diesem Zusammenhang: „Für die Überprüfung der Bildungsstandards wird ein Zeitpunkt ca. ein Jahr vor Abschluss des jeweiligen Bildungsgangs festgelegt. Auf der Grundlage repräsentativer Stichproben werden im Primarbereich in Jahrgangsstufe drei [...] zentrale Tests mit einem Ländervergleich durchgeführt [...]“ (KMK, 2006, S. 11). Dieses Vorgehen wird erst dadurch möglich, dass die Pilotierung und Normierung der entsprechenden Testaufgaben sowohl für Schülerinnen und Schüler der dritten als auch der vierten Jahrgangsstufe erfolgt ist.

Generell kann aus verschiedenen großen Schulleistungsstudien geschlussfolgert werden, dass je nach Fach und Jahrgangsstufe, die betrachtet werden, der Kompetenzzuwachs innerhalb eines Schuljahres zwischen einer viertel und einer halben Standardabweichung liegt (vgl. Bos et al., 2004). Hierbei fällt der Leistungszuwachs in unteren Jahrgangsstufen eher höher und in höheren Jahrgangsstufen eher geringer aus. Als Ursachen für diesen Befund kommen zum einen ein tatsächlich geringer Leistungszuwachs in höheren Jahrgangsstufen und zum anderen eine Vergrößerung der Leistungsstreuung im Verlauf der Schulausbildung in Betracht.

Bos et al. (2004, S. 56) berichten, dass der querschnittlich ermittelte Leistungsunterschied in der Lesekompetenz zwischen Dritt- und Viertklässlern in der Ergänzungsstudie IGLU-E 2001

43 Punkte beträgt. Setzt man diese Angabe mit den Ergebnissen für die vierte Klassenstufe (M = 539, SD = 67; Bos et al., 2003, S. 102) in Beziehung, so ergibt sich eine Effektstärke der Leistungsdifferenz von d = 0.64. Dieser Effekt liegt also in der Größenordnung von zwei drittel Standardabweichungen. Für die Bildungsstandards ergibt sich in der Pilotierungsstichprobe des Jahres 2006 als querschnittlich ermittelte Leistungsdifferenz der dritten und vierten Jahrgangsstufe eine exakte Replikation des Effekts von d = 0.64. Für die Normierungsstichprobe des Jahres 2007 finden wir eine Effektstärke von d = 0.58. Relevant ist in diesem Zusammenhang, dass diese Leistungsdifferenz nicht im Sinne einer direkten Messung der Leistungsentwicklung über die Zeit missverstanden werden darf, da keine längsschnittlichen Leistungszuwächse, sondern lediglich querschnittliche Beziehungen zwischen den Leistungen von Schülerinnen und Schülern der dritten und vierten Jahrgangsstufe berichtet werden. Nichtsdestotrotz stellen diese Befunde sehr gute Indikatoren des Mittelwertsverlaufes längsschnittlicher Leistungsentwicklung dar.

Von Bedeutung ist nun die Tatsache, dass eine deutliche Leistungsdifferenz zwischen Schülerinnen und Schülern des dritten und vierten Jahrgangs im Hinblick auf ihre Lesekompetenz zu erwarten ist. Die Formulierungen der Bildungsstandards im Fach Deutsch für den Primarbereich beziehen sich auf erwartete Kompetenzstände am Ende der vierten Jahrgangsstufe. Die Überprüfung einer prospektiv möglichen oder auch wahrscheinlichen Erreichung dieser Leistungserwartungen soll aber mehr als ein Schuljahr früher, nämlich im Verlauf der dritten Jahrgangsstufe, stattfinden. Aus diesem Umstand ergibt sich das Erfordernis, Testaufgaben zu entwickeln, die dem Kompetenzstand von Drittklässlern angemessen sind und gezwungenermaßen geringere Anforderungen stellen, als dies bei Aufgaben für die Überprüfung des Kompetenzstandes von Viertklässlern der Fall ist.

Konkret zeigte sich etwa, dass Aufgaben, bei denen Textsorten zu identifizieren waren, von Schülerinnen und Schülern der dritten Klasse oftmals im Sinne eines differenziellen Itemfunktionierens nicht gelöst werden konnten, da diese bestimmte Textsorten, beispielsweise Fabeln, noch nicht im Unterricht behandelt hatten. Die Bildungsstandards formulieren im Kompetenzbereich Lesen aber explizit die Erwartung, dass verschiedene Textsorten beherrscht und unterschieden werden können (vgl. KMK, 2005). Die Herausforderung besteht also darin, die Kompetenzstände von (bekanntermaßen) leistungsheterogenen Gruppen so zu erfassen, dass einerseits die minimalen und maximalen Leistungen beider Gruppen abgebildet werden und gleichzeitig eine hinreichend starke Verlinkung zwischen beiden Gruppen mittels identischer Items hergestellt werden kann. Anders formuliert ist es erforderlich, zwischen den (zumindest teilweise) distinkten Itemmengen, die der dritten und vierten Jahrgangsstufe vorgelegt werden, eine hinreichend starke Verbindung herstellen zu können. Dies ist notwendig, um in künftigen

Überprüfungen des Erreichens der Bildungsstandards auf der Grundlage einer in der dritten und vierten Jahrgangsstufe durchgeführten Normierung abschätzen zu können, ob die in der dritten Jahrgangsstufe getesteten Schülerinnen und Schüler die Standards werden erreichen können, wenn sie die vierte Jahrgangsstufe beendet haben werden.

Aufgrund der hier dargestellten Konstellation ergaben sich verschiedene designbedingte Probleme der Verlinkung der beiden interessierenden Jahrgangsstufen, die teils erhebliche Auswirkungen auf die resultierenden Schätzungen von Personenfähigkeiten und Itemschwierigkeiten hatten. Da diese methodischen Herausforderungen in der einschlägigen Literatur insbesondere im deutschen Sprachraum bislang wenig diskutiert wurden und somit eine breite Erfahrungsbasis fehlte, war es nur bedingt möglich, im Hinblick auf das verwendete Testdesign Entscheidungen zu treffen, die in jeder Hinsicht optimale Lösungen darstellen. Die resultierenden Hindernisse und Besonderheiten werden nachfolgend aufgezeigt und diskutiert.

IV.3.2 Methode

Charakterisierung der Personenstichproben

In den nachfolgend berichteten Analysen beziehen wir uns auf die Daten aus drei verschiedenen Studien, die in den Jahren 2006 bis 2008 im Rahmen der Pilotierung und Normierung der Bildungsstandards für den Primarbereich durchgeführt wurden. Tabelle IV.5 liefert einen zusammenfassenden Überblick der Charakterisierung der drei entsprechenden Teilstichproben, für die uns Daten zur Lesekompetenz vorliegen.

Tabelle IV.5: Stichprobencharakterisierung

Stichprobe	Testzeitraum	Stichprobenumfang		Anteil Mädchen in %	Alters-durchschnitt
		N (Schüler)	N (Klassen)		
Pilotierung	Frühjahr 2006	5585	318	48,8	9.9
Normierung I	Frühjahr 2007	3600	224	48,7	9.9
Normierung II	Frühjahr 2008	1683	80	48,0	9.9

Methodische Grundlagen der Verlinkung der Jahrgangsstufen drei und vier

Um eine Verbindung zwischen den Kompetenzständen der dritten und vierten Jahrgangsstufe herstellen zu können, muss in der einen oder anderen Form auf gemeinsame Itemmengen zurückgegriffen werden. Dies sind Items, die in beiden Jahrgängen zur Bearbeitung vorgelegt wurden und die aufgrund der Unterschiede in den Lösungshäufigkeiten eine Abschätzung der Leistungsdifferenz zwischen beiden Gruppen gestatten, die auf die distinkten Itemmengen generalisiert werden kann.

Grundsätzlich bieten sich hierfür entweder separate Skalierungen mit anschließendem Linking oder eine gemeinsame Skalierung der Daten beider Jahrgangsstufen an. Bei der gemeinsamen Skalierung beider Jahrgangsstufen kann wiederum ein Vorgehen mit und ohne Modellierung von Hintergrundinformationen gewählt werden. Bei einer gemeinsamen Skalierung beider Jahrgangsstufen „mit Hintergrundmodell" kann berücksichtigt werden, wie fähig beide Gruppen im Mittel sind.

An dieser Stelle soll aber insbesondere die vertikale Verlinkung näher betrachtet werden (vgl. Kolen & Brennan, 2004), da einparametrische Skalierungen mittels der Software ConQuest (vgl. Wu et al., 2007) bereits von Winkelmann und Böhme (2009) dargestellt wurden.

Die vertikale Verlinkung beider Jahrgangsstufen erreichten wir durch den Einsatz verschiedener Equating-Varianten. Allgemein sind Equating-Methoden lineare Transformationen, bei denen verschiedene Skalen möglichst „optimal" aufeinander abgebildet werden, indem Mittelwerte (und mitunter zusätzlich die Varianz) der Fähigkeitsverteilungen beider Gruppen berücksichtigt werden.

Die einfachsten Verfahren berücksichtigen hierbei nur die Mittelwerte sowie eventuell zusätzlich die Varianz von Itemparametern (Itemschwierigkeiten oder Itemtrennschärfen, falls ein zweiparametrisches Modell verwendet wird) aus IRT-Skalierungen (Mean-Mean-Equating, Mean-Sigma-Equating). Komplexere Equating-Methoden berücksichtigen zusätzlich weitere Merkmale der Fähigkeitsverteilung oder sind robuster hinsichtlich des „Vergleichs" der aus beiden Skalierungen gewonnenen Itemparameter. Die bekanntesten unter diesen komplexeren Prozeduren sind die *Characteristic Curve Methods* von Haebara (1980) und von Stocking und Lord (1983), welche die Item-Characteristic-Curves (d. h. die Item-Response-Funktion) der beiden aneinander anzugleichenden Skalen bedingt auf die Schülerfähigkeit θ „optimal verschieben".

Die beiden genannten Methoden unterscheiden sich im Wesentlichen darin, ob zuerst über die Items summiert oder zuerst die Differenz quadriert wird: Bei der Methode von Haebara (1980) wird für ein gegebenes θ die Differenz zwischen jeder Item-Characteristic-Curve der beiden Skalen quadriert und dann über die Items summiert. Bei Stocking und Lord (1983) wird dagegen zu einem gegebenen θ die quadrierte Differenz der Test-Characteristic-Curves

betrachtet, d. h. zunächst über die Itemfunktionen summiert und anschließend die Differenz gebildet und quadriert. Durch diese unterschiedlichen Vorgehensweisen können die Methoden unterschiedliche Ergebnisse erzeugen: Während sich bei Stocking & Lord gegenläufige Verzerrungen zwischen Items kompensieren können, zählt bei Haebara die Abweichung für jedes Item unabhängig von der Richtung dieser Abweichung. Bei der Methode von Stocking & Lord wird demzufolge in der Suche nach der Verschiebungskonstanten beim Equating ein differenzielles Itemfunktionieren (*Item-DIF*) zwischen beiden Stichproben zugelassen, da alle Itemfunktionen zur *Test-Characteristic-Curve* addiert werden. Genauere Details zu diesen Verfahren findet man in einem Überblick in Yen und Fitzpatrick (2006, S. 133ff.) oder Kolen und Brennan (2004).

Differential Item Functioning und Differential Testlet Functioning (DIF, DTF)

Differenzielles Itemfunktionieren (*Differential Item Functioning,* DIF) ist ein zentrales Kriterium bei der Beurteilung der Testfairness (vgl. AERA, APA & NCME, 1999). Formal definiert liegt bei einem Item dann DIF vor, wenn sich Testteilnehmer, die unterschiedlichen Subgruppen der Population entstammen, aber im Hinblick auf das mit dem Test gemessene Konstrukt gleiche Fähigkeiten aufweisen, in der Beantwortung des Items unterscheiden (vgl. Penfield & Algina, 2006; vgl. Roussos & Stout, 2004).

Für gewöhnlich werden beispielweise Subgruppen nach Geschlecht oder Migrationshintergrund unterschieden. Es ist aber auch möglich, das differenzielle Itemfunktionieren hinsichtlich verschiedener Jahrgangsstufen zu untersuchen. Vereinfacht bedeutet die Existenz von Klassenstufen-DIF, dass – *nach* Kontrolle eines bestehenden Fähigkeitsunterschiedes zwischen den Jahrgangsstufen – manche Items eher Schüler der dritten Klassenstufe bevorzugen, andere Items wiederum tendenziell Schüler der vierten Klassenstufe. Mit „bevorzugen“ meinen wir hierbei, dass den jeweiligen Schülerinnen und Schülern die Lösung der entsprechenden Items leichter fällt, als dies – gegeben ihre Fähigkeiten in Übereinstimmung mit dem Modell – der Fall sein sollte.

Definitionsgemäß beträgt der Mittelwert des Klassenstufen-DIFs über alle Items hinweg null. Longford, Holland und Thayer (1993) sowie Camilli und Penfield (1997) argumentieren daher, dass die Varianz der DIF-Werte (im Folgenden auch als DIF-Varianz oder DIF-Standardabweichung für die Wurzel aus der korrespondierenden Größe bezeichnet) eine interessante Skaleneigenschaft darstellt. Bezogen auf DIF-Effekte zwischen den Jahrgangsstufen gibt die DIF-Varianz das Ausmaß der Heterogenität des Itemfunktionierens zwischen den Klassen drei und vier an. Bei einer Varianz von null würde keine Heterogenität und damit kein

Klassenstufen-DIF auf den Items existieren. In diesem wünschenswerten Fall wäre die Wahl einer Menge von Verlinkungsitems allein durch die Passung ihrer Schwierigkeit auf beide Personenpopulationen begründet. Existiert jedoch DIF-Varianz und wählt man eine Teilmenge von Items als Verlinkungsitems, so besteht die Gefahr einer Verschätzung des mittleren Unterschieds zwischen beiden Populationen. Eine Verschätzung des Leistungsunterschieds zwischen den Jahrgangsstufen tritt dann ein, wenn alle Verlinkungsitems (tendenziell) DIF in eine bestimmte Richtung aufweisen, und zwar im Umfang des mittleren DIF-Effekts der Verlinkungsitems.

DIF-Varianz induziert somit einen zusätzlichen Fehler bei der Verlinkung von Schülerinnen und Schülern der dritten und vierten Jahrgangsstufe auf einer Skala (einen so genannten *Linking Error*, vgl. Monseur & Berezner, 2007), sodass die Erfassung von Leistungsunterschieden mit einer größeren Unsicherheit behaftet ist. Allerdings wirkt sich DIF-Varianz nur dann auf den Linkingfehler aus, wenn die Verlinkung verschiedener Jahrgangsstufen mittels einer Itemstichprobe vorgenommen wird, die aus einem größeren (endlichen, aber als unendlich denkbaren) Itempool stammt. Für einen Test mit feststehenden Items liegt selbst bei einer stark ausgeprägten Existenz von DIF-Varianz kein Linking Error vor, da kein Item Sampling stattfindet und demzufolge die Leistungsunterschiede zwischen den Jahrgangsstufen mit genau den im Test enthaltenen Items erfasst werden.

Tabelle IV.6 gibt als Interpretationshilfe einen Überblick, in welchen Größenordnungen DIF und DIF-Varianz als kritisch beurteilt werden müssen. Diese Angaben beziehen sich wesentlich auf die von Penfield und Algina (2006) vorgestellten Richtlinien für die Interpretation der DIF-Varianz (Penfield & Algina, 2006, S. 307f.), sowie auf die von Monahan, McHorney, Stump und Perkins (2007) ausgearbeitete Gegenüberstellung verschiedener DIF-Kennwerte.

Tabelle IV.6: In der Literatur berichtete Interpretationen von DIF-Kenngrößen

DIF Item Effekt (ETS-Klassifikation)			
Wertebereich		Interpretation	Quelle
> .64 und signifikant > .43		großer DIF-Effekt (*C DIF Item*)	Monahan u.a. (2007)
.43 – .64 und signifikant > 0		moderater DIF-Effekt (*B DIF Item*)	
< .43		vernachlässigbarer DIF-Effekt (*A DIF Item*)	
DIF Variance			
Wertebereich (DIF-Varianz)	Wertebereich (DIF-Standard-abweichung)	Interpretation	Quelle
> .14	> .37	großer Effekt	Penfield & Algina (2006)
.07 – .14	.26 – .37	mittlerer Effekt	
< .07	< .26	kleiner Effekt	

Bei Items zur Erfassung der Lesekompetenz kann nun zusätzlich zu einem differenziellen Funktionieren auf Itemebene so genanntes *Differential Testlet Functioning* (DTF; vgl. Wainer, Sireci & Thisse, 1991) auftreten. Hierbei umfasst ein Testlet jeweils einen Lesetext und alle zu diesem Text gehörenden Items. Testlet-DIF berücksichtigt, dass Items, die sich auf ein und denselben Lesetext beziehen, einander ähnlicher sind als Items, die zu unterschiedlichen Stimulustexten gestellt werden.

Im klassischen Sinn wird ein Testleteffekt als Methodeneffekt verstanden, der Varianz zwischen Schülerleistungen verursacht (vgl. Robitzsch, 2009). Da man davon ausgehen kann, dass ein bestimmter Text einigen Schülern beispielsweise inhaltlich mehr entgegenkommt als anderen Schülern, kann es bei der Bearbeitung von Leseaufgaben zu einer „Interaktion“ von Schülern mit den Items jeweils eines Lesetextes kommen. In Testlet-Modellen (vgl. Bradlow, Wainer & Wang, 1999; vgl. Wainer, Bradlow & Wang, 2007) wird daher neben der primär zu erfassenden Kompetenz zusätzlich für jedes Testlet und jeden Schüler ein spezifischer Faktor eingeführt, der als Methodeneffekt für den jeweiligen Lesetext interpretiert werden kann und der die stochastische Abhängigkeit zwischen den Items zu einem gemeinsamen Text quantifiziert. Testlet-DIF hingegen kennzeichnet den mittleren Item-DIF über alle Itemschwierigkeiten eines

Stimulustextes hinweg. Während also Testleteffekte das Ausmaß an durch Sekundärdimensionen verursachter Varianz zwischen Schülerleistungen quantifizieren, setzt Differential Testlet Functioning am differenziellen Itemfunktionieren (und damit verschiedenen Itemschwierigkeiten) an und untersucht, ob diese Effekte durch den Faktor des gemeinsamen Testlets erklärt werden können.

Die Variabilität dieses DIF-Wertes in der Stichprobe aller Texte ist insofern relevant, als für Verlinkungen keine einzelnen Items, sondern immer nur Itembündel gewählt werden können, nämlich die zu einem Lesetext gehörenden Items. Zur Erfassung des Testlet-DIFs zerlegen wir deshalb den beobachteten DIF-Wert für jedes Item eines Textes in zwei Varianzanteile: in den Testlet-DIF-Effekt und in den residualen Itemeffekt. In der Terminologie der Mehrebenenmodelle stellen dabei die Items die Level-1-Einheiten, Aufgaben oder Texte die Level-2-Einheiten dar. Diese beiden Varianzquellen wurden nach Korrektur der Messfehler für die Items aller Verlinkungsaufgaben der Normierungsstudie in WinBUGS (Spiegelhalter, Thomas, Best & Lunn, 2003) ermittelt. Hierbei handelt es sich um 16 Leseaufgaben, die als Stimuli literarische Texte sowie kontinuierliche und diskontinuierliche Sachtexte verwenden. Diese Aufgaben umfassen insgesamt 90 Items unterschiedlicher Formate.

IV.3.3 Befunde zur Verlinkung der Jahrgangsstufen drei und vier sowie zum differenziellen Itemfunktionieren

Befunde zur Verlinkung beider Jahrgangsstufen

Für eine vertikale Verlinkung beider Jahrgänge wurden die Daten der Klassenstufen drei und vier zunächst getrennt mithilfe der Software ConQuest (vgl. Wu et al., 2007) skaliert. In diese Skalierungen wurden die Stichproben der Dritt- beziehungsweise Viertklässler aus der Pilotierungsstudie 2006, der Normierungsstudie 2007 sowie einer zweiten Normierungsstudie aus dem Jahr 2008 einbezogen. Diese beiden getrennten Skalierungen für die Klassenstufen drei und vier wurden mittels Equating miteinander in Beziehung gesetzt. Die verschiedenen, oben diskutierten Equating-Methoden liefern im vorliegenden Fall für die vertikale Verlinkung der dritten und vierten Jahrgangsstufe allerdings sehr ähnliche Schätzungen für die Verschiebungskonstante. Diese ermittelte Verschiebungskonstante ist als Fähigkeitsunterschied zwischen Dritt- und Viertklässlern für die Itemstichprobe der Verlinkungsitems in Einheiten der Logitskala interpretierbar. Es resultiert sowohl für das Mean-Mean-Equating als auch für die Equating-Varianten von Stocking und Lord und von Haebara eine Logitdifferenz von .67 bis .68.

Zwischen den Ergebnissen der erläuterten Methoden zeigt sich also kein praktisch relevanter Unterschied.

Um die Parameter der Skalierung aus der dritten Jahrgangsstufe auf die Metrik der vierten Jahrgangsstufe zu überführen und beide Jahrgangsstufen innerhalb eines gemeinsamen Kompetenzstufenmodells verorten zu können, wurde für Fähigkeitsschätzungen und Itemparameter der dritten Jahrgangsstufe die ermittelte Verschiebungskonstante addiert. Auf diese Weise wurde eine gemeinsame Metrik der dritten und vierten Jahrgangsstufe etabliert.

Befunde zu verschiedenen Methoden der vertikalen Skalierung

Wie bereits angedeutet, ist die gemeinsame Skalierung beider Jahrgänge ein alternativer Zugang, um eine Abbildung beider Klassenstufen auf einer einheitlichen Skala zu erreichen. Hierbei kann eine Skalierung ohne Hintergrundmodell und damit ohne explizite Information darüber, welcher Jahrgangsstufe ein Schüler angehört, von einer Skalierung mit Hintergrundmodell, bei der für jeden Schüler die Gruppenzugehörigkeit explizit definiert wird, unterschieden werden.

In Abbildung IV.3 sind die Itemparameterschätzungen aus einer gemeinsamen Skalierung ohne Berücksichtigung der verschiedenen Leistungsstände und somit der verschiedenen Gruppenmittelwerte („ohne Gruppe“) im Vergleich zur Skalierung mit einem Hintergrundmodell für die Jahrgangsstufen („mit Gruppe“) dargestellt. Man erkennt, dass die Abweichung in der Schätzung der Itemparameter davon abhängt, in welcher der beiden Jahrgangsstufen das jeweilige Item eingesetzt wurde. Ausschließlich in Klassenstufe drei eingesetzte und damit tendenziell leichtere Items weisen erhöhte Schwierigkeitsschätzungen bei der Analyse „ohne Gruppe“ auf. Die Items werden also als schwerer eingestuft, als sie eigentlich sind. Der Unterschied zwischen den beiden Skalierungsvarianten beträgt für diese Items etwa –.50 Logits, was praktisch relevant und somit nicht vernachlässigbar ist. Umgekehrt werden die relativ schwereren Items aus Klassenstufe vier bei der Skalierung ohne Berücksichtigung der Gruppenstruktur um etwa .50 Logits unterschätzt, sodass die Items als leichter beurteilt werden, als sie tatsächlich sind. Bei Verlinkungsitems, die sowohl von Schülerinnen und Schülern der dritten als auch der vierten Jahrgangsstufe bearbeitet wurden, existiert kein praktisch bedeutsamer Unterschied zwischen den Skalierungsvarianten. Zusammenfassend replizieren die in Abbildung IV.3 dargestellten Ergebnisse den wesentlichen Befund von DeMars (2002).

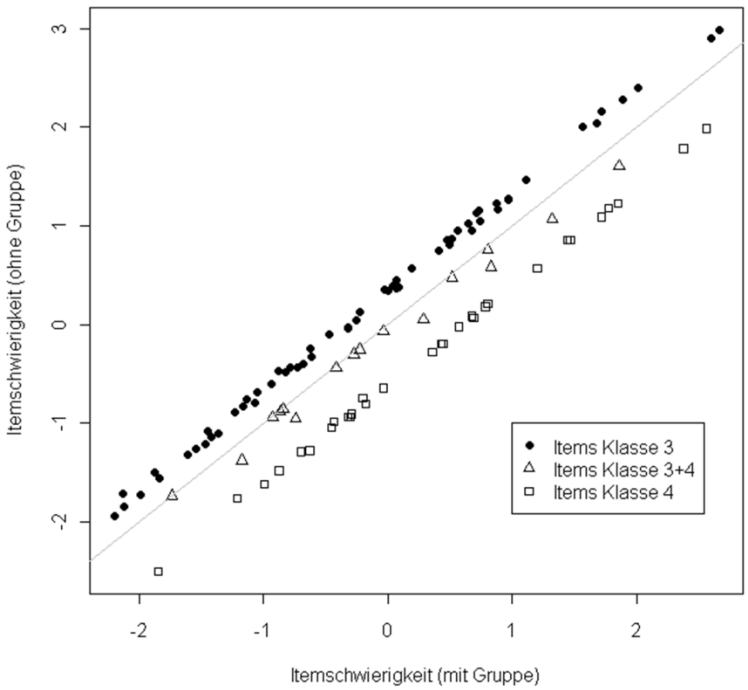

Abbildung IV.3: Vergleich der Itemparameterschätzungen in einer gemeinsamen Skalierung für die dritte und vierte Jahrgangsstufe mit und ohne Hintergrundmodell für die Gruppenzugehörigkeit

Die von Robitzsch (2009) vorgestellten Befunde zum Vergleich verschiedener Equatingverfahren werden somit auch anhand der hier erläuterten Ergebnisse für reale Daten bestätigt. Bei einer gemeinsamen Skalierung unter Hinzunahme eines Hintergrundmodells für die Gruppenzugehörigkeit fällt die Schwierigkeit leichter Items im Mittel etwas geringer aus als mit der Methode des Equatings für eine vertikale Verlinkung. Aufgrund der Simulation ist aber davon auszugehen, dass es sich bei dem vorliegenden Effekt von kleiner als –.10 Logits um eine (praktisch wenig relevante) Verzerrung der Itemschwierigkeitsschätzung handelt. Der umgekehrte Effekt wird für die in Jahrgangsstufe vier eingesetzten Items beobachtet.

Insgesamt kann festgehalten werden, dass sich im Vergleich der Methoden – der Skalierung mit Hintergrundmodell für die Gruppenzugehörigkeit einerseits und dem Equating andererseits – sehr ähnliche Itemparameterschätzungen ergeben.

Einschränkend sollte festgehalten werden, dass in der gemeinsamen Skalierung beider Jahrgangsstufen in ConQuest die Annahme gleicher Varianzen der Fähigkeitsverteilungen in den Klassenstufen drei und vier getroffen wird. Diese ist im Fall der Lesekompetenz jedoch stark verletzt (Varianz Klassenstufe drei: 1.58; Varianz Klassenstufe vier: 1.01). Die Restriktion gleicher

Varianzen oder die Annahme von Normalverteilungen kann bei Verwendung alternativer IRT-Software (ICL: vgl. Hanson, 2002, oder Mplus: vgl. Muthén & Muthén, 2006) umgangen werden.

Da die Ergebnisse von Robitzsch (2009) eine Abhängigkeit der Unterschiede zwischen den Skalierungsvarianten, bedingt durch die Stärke der Verlinkung innerhalb des Testdesigns, nahezulegen scheinen, gehen wir infolge der Geringfügigkeit der Unterschiede zwischen der Skalierung mit Hintergrundmodell einerseits und einem Equating andererseits von einem hinreichend verlinkten Design aus.

Befunde zum differenziellen Itemfunktionieren zwischen den Klassenstufen

Wie bereits erläutert, muss für die Verlinkung der Jahrgänge drei und vier eine Teilmenge an Items aus dem gesamten Itempool in beiden Klassenstufen vorgelegt werden. Ebenfalls diskutiert wurde die Tatsache, dass es sich bei den Schülerinnen und Schülern dieser beiden Jahrgänge um leistungsheterogene Gruppen handelt, da Drittklässler im Mittel geringere Kompetenzstände aufweisen als Viertklässler. Zur Abbildung dieser Leistungsunterschiede müssen die zur Verlinkung eingesetzten Items nun einerseits sensitiv gegenüber diesen querschnittlichen Leistungsdifferenzen sein. Andererseits müssen sie gleichzeitig der Forderung genügen, als Itemgruppe möglichst homogen zu funktionieren. Vor diesem Hintergrund ergibt sich die Frage, inwieweit es zwischen den Klassenstufen drei und vier zu differenziellem Funktionieren von Items und Aufgaben kommt.

Zunächst wurde für die Pilotierungs- und Normierungsstichprobe auf der Ebene der Items der Klassenstufen-DIF für alle Verlinkungsitems bestimmt. Die durch IRT-Skalierungen geschätzten Itemparameter sind messfehlerbehaftet. Die Größe des Messfehlers ist jedoch durch die geschätzten Standardfehler gegeben, woraus man nach Elimination dieses Messfehlers die DIF-Varianz berechnen kann. Dies ist im Rahmen von *Linear Mixed Models* mit bekannter Level-1-Varianz (*Variance Known Models*; Raudenbush & Bryk, 2002, S. 205ff.) möglich. Mithilfe der Software WinBUGS (vgl. Spiegelhalter et al., 2003) durchgeführte Analysen ergaben für die Normierungsstudie (2007) eine Standardabweichung des Klassenstufen-DIFs von .31, für die Pilotierungsstudie (2006) von .32, sodass trotz verschiedener Verlinkungsitems in beiden Personenstichproben bezüglich des DIF-Effekts von einer stabilen Skaleneigenschaft ausgegangen werden kann. Für die Verlinkungsitems der Normierungsstudie bedeutet dies unter Berücksichtigung der Normalverteilungsannahme, dass etwa 68 % der Items DIF-Werte zwischen –.31 und .31 aufweisen, während 32 % der Items betraglich größere DIF-Werte als .31 zeigen. Dies entspricht einer DIF-Varianz .10, was gemäß Tabelle IV.5 als mittlerer Effekt interpretiert werden kann.

In einem zweiten Schritt wurde das differenzielle Funktionieren auf der Ebene von Aufgaben beziehungsweise Stimuli geprüft. Wie oben erläutert kann ein Lesetext mit den zugehörigen Items als ein Testlet verstanden werden, da sich die Items eines Textes untereinander ähnlicher sind als im Vergleich zu Items eines anderen Stimulus. Die Standardabweichung des DIF auf Aufgabenebene beträgt .30 (81.6 % der DIF- Varianz), während die DIF-Standardabweichung der Items innerhalb von Aufgaben nur .14 beträgt (18.4 % der DIF-Varianz). Konkret variieren die Schätzungen der Testlet-DIF-Effekte (Mittelwerte der Posteriorverteilungen; siehe Wang, Bradlow, Wainer & Muller, 2007) für die 16 Testlets, also Leseaufgaben, zwischen –.39 und .46 Logits. Demzufolge ist der größere Varianzanteil des klassenstufenspezifischen Itemfunktionierens auf Aufgabenebene zu finden. Ein positiver DIF-Wert bedeutet hierbei, dass das entsprechende Item beziehungsweise die entsprechende Aufgabe für einen Drittklässler schwieriger ausfällt, als gemäß dem Rasch-Modell zu erwarten ist. Drittklässler werden also benachteiligt, Viertklässler bevorzugt. Dementsprechend bedeutet ein negativer DIF-Wert, dass Drittklässler bevorzugt beziehungsweise Viertklässler benachteiligt werden.

Die Aufgabe, welche die größte Bevorzugung der dritten Jahrgangsstufe aufweist, verwendet als Stimulus kurze, inhaltlich voneinander unabhängige Aussagen, die bestimmten Kategorien zugeordnet werden müssen. Dieser Stimulus greift stärker als andere Texte auf das Vorwissen der Schülerinnen und Schüler zurück, da die fraglichen Kategorien als bekannt vorausgesetzt werden. Eine Zuordnung des Stimulus entsprechend der üblichen Einteilung in Sachtexte vs. literarische Texte beziehungsweise kontinuierliche vs. diskontinuierliche Texte ist aufgrund seiner besonderen Beschaffenheit nur schwer möglich. Der verwendete Stimulus ist kein prototypischer Vertreter der diskutierten Kategorien, daher scheint eine Verallgemeinerung dahingehend, dass mit dieser Stimulusart generell eine Bevorzugung der dritten Jahrgangsstufe einhergeht, unzulässig.

Die drei Aufgaben, die eine recht deutliche Bevorzugung der vierten Klassenstufe zeigen, verwenden als Stimulus einen kontinuierlichen und zwei diskontinuierliche Sachtexte. Bei beiden diskontinuierlichen Texten werden Informationen in einer Tabelle dargeboten. Diese Tabellen weisen im einen Fall recht komplexe Schachtelungen auf und sind in beiden Fällen sehr umfangreich (mindestens drei Spalten, mehr als zehn Zeilen). Es scheint daher plausibel, dass Drittklässler mit dieser Art der Informationsdarstellung erstens weit weniger vertraut und zweitens möglicherweise kognitiv überfordert sind. Es liegt die Vermutung nahe, dass mit der Wahl eines diskontinuierlichen Sachtextes in Form einer komplexen Tabelle generell eine Benachteiligung der Drittklässler (über den zu erwartenden Leistungsunterschied der dritten und vierten Jahrgangsstufe hinaus) einhergehen könnte.

Der kontinuierliche Sachtext beschäftigt sich mit der Lebensweise einer Vogelart. Das Sprachniveau ist wenig komplex, der Lesbarkeitsindex LIX nach Björnsson (vgl. Köster, 2005) ebenso wie die vierte Wiener Sachtextformel weisen den Stimulustext als einfach aus. Schwierig und an dem Modell gemessen „zu schwierig" für die dritte Jahrgangsstufe wird der Text dadurch, dass er im Testdesign gemeinsam mit einem zweiten Stimulus in einem Block dargeboten wird, wobei allerdings keine Bezugnahme gefordert wird. Die kognitive Anforderung, beide Aufgabenstellungen direkt nacheinander bearbeiten zu müssen, scheint aber für Drittklässler unverhältnismäßig schwieriger zu sein. Daher sollte für die Textgattung der kontinuierlichen Sachtexte keine Verallgemeinerung einer generellen Benachteiligung von Drittklässlern getroffen werden.

Weiterhin fällt auf, dass unter denjenigen Aufgaben, die einen großen Testlet-DIF-Effekt (betraglich > .30 Logits) aufweisen, ausschließlich kontinuierliche und diskontinuierliche Sachtexte zu finden sind. Alle kontinuierlichen literarischen Texte zeigen Testlet-DIF-Effekte lediglich in einem Umfang, der praktisch nicht bedeutsam ist.

Einschränkend muss allerdings berücksichtigt werden, dass diese DIF-Effekte auf Aufgabenebene mit Positions- und Kontexteffekten der Items konfundiert sind, da nicht alle Blockpositionen der Aufgaben in beiden Stichproben vollständig variiert werden konnten (vgl. Erläuterungen zum Testdesign in Winkelmann & Böhme, 2009).

Nachfolgend wird argumentiert, dass auch nach Kontrolle dieser Effekte noch immer ein substanzieller Varianzanteil des Klassenstufen-DIFs auf Aufgabenebene verbleiben dürfte. So wurde zum Beispiel in den internationalen PISA-Studien 2000 und 2003 für die in beiden Jahren eingesetzten Ankeritems geprüft, ob eine Veränderung der Itemparameter im Sinne eines differenziellen Funktionierens der Items zwischen beiden Erhebungen eingetreten ist. Es wurden DIF-Effekte ermittelt, die auf ein verschiedenes Funktionieren der Items in den Jahren 2000 und 2003 hinweisen. Für die Ankeritems wurde eine DIF-Varianz von .05 berichtet (Monseur & Berezner, 2007, S. 326). Vor dem Hintergrund dieses Befundes kann nun die in unseren Studien für das Leseverstehen zwischen den Jahrgangsstufen drei und vier beobachtete DIF-Varianz von .09 bei variierenden Populationen wie folgt diskutiert werden: In PISA wurden über die Studien hinweg unveränderte Blöcke aus Leseaufgaben in äquivalenten Schülerpopulationen eingesetzt. Somit sind lediglich Positions- und Kontexteffekte zu erwarten, da die Leseblöcke in den Testheften an unterschiedlichen Positionen gemeinsam mit Aufgaben verschiedener Domänen (Lesen, Mathematik und Naturwissenschaften) zur Bearbeitung vorgelegt wurden. Unter Rückgriff auf diese Ergebnisse könnte man nun argumentieren, dass sich als wahre Varianz des Klassenstufen-DIFs die Differenz aus beobachteter Klassenstufen-DIF-Varianz in unserer Studie und der designbedingten DIF-Varianz in PISA ergeben könnte. Eine entsprechende

Überschlagsrechnung liefert als „wahre DIF-Varianz“ für den Klassenstufen-DIF .09 – .05 = .04. Diese hypothetische Schätzung des wahren Effekts bedarf natürlich einer Replikation in weiteren Studien.

Auch die Testlet-DIF-Varianz wurde im Zuge des Vergleichs der PISA-Studien 2000 und 2003 untersucht. Dabei entfielen etwa 50 % der gesamten DIF-Varianz der Items auf die Testletebene. Diese Größenordnung ist vorrangig auf Positions- und Kontexteffekte zurückzuführen. Demzufolge wird vermutlich auch in unserer Studie ein beträchtlicher Varianzanteil auf Positions- und Kontexteffekte zurückführbar sein. Der in unserer Studie ermittelte Varianzanteil auf Testletebene ist jedoch mit etwa 80 % noch deutlich größer als der in PISA ermittelte Anteil. Führen wir auch hier eine Überschlagsrechnung als Vergleich zwischen der PISA- und unserer Studie durch, so ergibt sich immer noch ein Anteil von 80 – 50 = 30 % wahrer Testlet-Varianz auf Klassenstufenebene, der wie oben ausgeführt qualitativ interpretiert werden kann.

IV.3.4 Zusammenfassung und Diskussion

Die vorgestellten Ergebnisse (vgl. „Befunde zur Verlinkung beider Jahrgangsstufen“ und „Befunde zu verschiedenen Methoden der vertikalen Skalierung“) zu den ermittelten Leistungsunterschieden zwischen den Jahrgangsstufen drei und vier zeigen, dass sich die aus der Literatur bekannten Differenzen in unserer Studie weitgehend replizieren lassen. Von primärem Interesse war für uns jedoch die Frage, welche Konsequenzen verschiedene methodische Varianten der Verlinkung mit sich bringen. Hierbei zeigte sich, dass innerhalb der vertikalen Verlinkung mittels Equating verschiedene Ansätze zu nahezu identischen Ergebnissen führen und als mittlere Differenz der Verlinkungsitems zwischen den Jahrgangsstufen drei und vier eine Logitdifferenz von .67 bis .68 resultiert. Wählt man statt eines Equating-Zugangs eine gemeinsame Skalierung der zu verlinkenden Gruppen, so ergeben sich deutliche Unterschiede in Abhängigkeit davon, ob in die Modellierung Informationen über die Gruppenzugehörigkeit und somit über die unterschiedlichen Kompetenzstände dieser Gruppen einfließen oder nicht. Werden in einem Modell ohne Hintergrundinformation über die mittleren Fähigkeiten der Gruppen die Schwierigkeiten von Items geschätzt, die ausschließlich einer der beiden Gruppen zur Bearbeitung vorgelegt wurden, so können sich Verschätzungen in der Größenordnung eines halben Logits ergeben. Es sollte daher (bei moderater Verlinkung durch identische Aufgaben und Items) entweder ein Equating oder eine Skalierung mit Hintergrundmodell gewählt werden, wobei die Ergebnisse in der Literatur hinsichtlich einer Empfehlung, welche Methode zu bevorzugen ist, differieren. Die Simulationsergebnisse von Robitzsch (2009) sprechen jedoch

eher für separate Skalierungen. Von entscheidender Bedeutung ist jedoch die Verlinkung beider Klassenstufen durch die Berücksichtigung hinreichend vieler gemeinsamer Aufgaben und Items. Im Verhältnis zur endlichen oder unendlichen angestrebten Itempopulation sollte dabei die Menge der jeweils distinkten Items, die also nur von Dritt- oder nur von Viertklässlern bearbeitet werden, möglichst klein gehalten werden.

Im Hinblick auf das differenzielle Itemfunktionieren zwischen dritter und vierter Jahrgangsstufe veranschaulichen die präsentierten Befunde, dass das Ausmaß des Klassenstufen-DIFs zwischen den beiden Jahrgängen insgesamt eher moderat ausfällt. Zerlegt man die Effekte in einen Testlet- und einen Itemanteil, so ist der deutlich größere Effekt auf das Testlet, also spezifische Aufgaben beziehungsweise Stimuli, zurückzuführen. Möglicherweise führen diskontinuierliche Sachtexte in Form komplexer Tabellen generell zu einer Benachteiligung von Drittklässlern. Diese Analysen könnten im Rahmen hierarchisch logistischer Regressionsmodelle zur Erklärung von DIF durch Kovariaten auf Item- und Stimulusebene angereichert werden (vgl. Swanson, Clauser, Case, Nungester & Featherman, 2002). DIF-Untersuchungen könnten somit einerseits Hypothesen generierend, andererseits in nachfolgenden Studien auch Hypothesen prüfend eingesetzt werden.

IV.4 Zusammenführung der Befunde und Ableitung von Empfehlungen

Im vorliegenden Beitrag wurden zwei methodische Fragestellungen thematisiert, die uns bei der Erfassung der Lesekompetenz beschäftigten. Zum einen interessierte uns die Untersuchung der Konstruktdimensionalität der Lesekompetenz und zum anderen die Betrachtung des differenziellen Itemfunktionierens in leistungsheterogenen Gruppen.

Die präsentierten Ergebnisse zur Konstruktdimensionalität der Lesekompetenz im Primarbereich wurden dahingehend interpretiert, dass zwar keine klar eindimensionale Struktur identifiziert werden konnte, aber auch keine eindeutig mehrdimensionale Struktur zu erkennen ist, die eine theoretisch plausible Differenzierung verschiedener Subfacetten gestatten würde.

Dies legt nahe, für die Beschreibung der Kompetenzstände der Schülerinnen und Schüler keine getrennten Kompetenzstufenmodelle für verschiedene Facetten der Lesekompetenz, wie etwa für das Verstehen kontinuierlicher literarischer Texte, zu wählen. Vielmehr erscheint ein einheitliches Kompetenzstufenmodell sinnvoll, wobei – ähnlich wie in IGLU (vgl. Bos et al., 2007) – auf den einzelnen Stufen mitunter Hinweise auf Besonderheiten der Verstehensleistungen für einzelne Textsorten zu geben sind. Separate Stufenmodelle für

verschiedene Subdimensionen der Lesekompetenz scheinen uns allerdings nicht angezeigt. Allgemein sollte die Klärung der dimensionalen Struktur eines Testkonstrukts stets von der Überlegung geleitet werden, dass Sammlungen an Testitems äußerst selten eine einzige Fähigkeit messen, im Normalfall wird ein Komplex von Fähigkeiten angesprochen (vgl. Reckase et al., 1988). Mehrdimensionalität ist daher oftmals eine Frage des interessierenden Auflösungsgrades (*Granularity*) des zu messenden Konstrukts. Ob Konstruktmerkmale als schwierigkeitsbeeinflussend für Items (vgl. Gorin & Embretson, 2006) oder als hochdimensionale (gegebenenfalls diskrete latente) Personeneigenschaften (*Cognitive Diagnostic Models*; DiBello, Roussos & Stout, 2007; Rupp & Templin, 2008) aufgefasst werden, hängt – neben modelltheoretischen Präferenzen – vom Zweck einer statistischen Modellierung ab.

Ein Forschungsdesiderat, welches sich aus den oben berichteten Befunden zu der in DETECT explorativ ermittelten Itemclusterung ergibt, wäre die inhaltlich-didaktische Analyse dieser Gruppierungen und die Frage, ob sich hieraus möglicherweise alternative, plausibel interpretierbare Dimensionen der Lesekompetenz im Primarbereich ergeben.

Douglas, Roussos und Stout (1996) schlagen vor, mittels DETECT identifizierte Teilmengen von Items (Itembündel) hinsichtlich DIF zu untersuchen. Hinter diesem Ansatz steht die Vorstellung, dass DIF auf Itemebene aufgrund verschiedener Fähigkeitsausprägungen in unterschiedlichen Personengruppen auf einer Sekundärdimension zustande kommt (vgl. Roussos & Stout, 1996). Mithilfe einer Abweichung von (essenzieller) Eindimensionalität wird daher modellkonträres Verhalten erklärt, sodass dieses Vorgehen Hypothesen geleitet mit Untersuchungen der Dimensionalität verbunden ist: „[...] an intelligently conducted DIF analysis must go hand-in-hand with a careful consideration of multidimensionality“ (Douglas et al., 1996, S. 482).

Anhand der vorgestellten Interpretation des Klassenstufen-DIFs am Beispiel der Lesekompetenz wird dieses Erklärungsmuster bestätigt. Gemäß der Definition des mehrdimensionalen DIF-Paradigmas (vgl. Shealy & Stout, 1993) unterscheidet man die den DIF verursachenden sekundären Fähigkeiten (*Abilities*) in *Nuissance Abilities* und *Auxiliary Abilities*. Die Nuissance Abilities („Stördimension“) messen konstruktirrelevante Fähigkeiten (etwa benötigtes oder vorteilhaftes Hintergrundwissen für spezifische Lesetexte) und werden als „wahrer“ zu vermeidender DIF angesehen. Davon sind Auxiliary Abilities als konstruktinhärente Dimensionen zu unterscheiden, die zwar DIF verursachen, jedoch Bestandteil des Konstrukts sind (etwa bei verschiedenen Textsorten), bei denen unterschiedliche Leistungen von Subpopulationen zugelassen werden müssen. Differenzielles Item- oder Testletfunktionieren im Zusammenhang mit Auxiliary Abilities ist damit per Definition *kein* Itemselektionskriterium.

In Bezug auf die Verlinkung der leistungsheterogenen Jahrgangsstufen drei und vier ergab sich, dass eine hinreichend dichte Verlinkung unbedingte Voraussetzung für die Herstellung eines Zusammenhangs zwischen diesen Gruppen ist. Die Wahl der Analysemethode zur Verlinkung (Equating, Skalierung mit Hintergrundmodell) ist hierbei eher sekundär, solange eine ausreichende Überlappung der für die betreffenden Gruppen ausgewählten Itemmengen im Design vorgesehen ist. Hierbei sollte berücksichtigt werden, dass bei der Erfassung der Lesekompetenz Items in Aufgaben geschachtelt sind und DIF sowohl auf Item- als auch auf Aufgabenebene zutage treten kann. Zeigt eine bestimmte Aufgabe auf Ebene des Testlets eine spezifische Bevorzugung oder Benachteiligung, so erweisen sich möglicherweise nicht nur ein oder zwei Items für die Verlinkung als unbrauchbar, sondern eine ganze Aufgabe. Das Maß einer hinreichenden Verlinkung sollte daher nicht nur die absolute Anzahl der Verlinkungsitems sein, sondern auch die Anzahl der Verlinkungsaufgaben, da Verlinkungsfehler aus den beiden Varianzkomponenten der Aufgaben- und Itemebene entstehen (Monseur & Berezner, 2007, S. 327). Eine Einschränkung der Aufgabenanzahl führt so möglicherweise zu einer Einschränkung der Validität von Aussagen über Klassenstufenunterschiede.

Die bildungspolitischen Vorgaben legen in diesem Zusammenhang nahe, solche Aufgaben und Items zu entwickeln, die spezifisch die jeweiligen Kompetenzstände beider Jahrgangsstufen abbilden. Methodische Beschränkungen und die oben vorgestellten Befunde verlangen jedoch eine möglichst dichte Verlinkung beider Jahrgänge über identische Aufgaben und Items. Den Schülerinnen und Schülern spezifische und damit distinkte Aufgaben- und Itemmengen zur Bearbeitung vorzulegen, ist kein gangbarer Weg. Vielmehr sollte das zukünftige Ziel eine Orientierung an änderungssensitiven Aufgaben und Items sein, die Lernzuwächse adäquat abbilden können und für beide Jahrgangsstufen in gleicher Weise geeignet sind, ohne zu einer spezifischen Bevorzugung oder Benachteiligung des einen oder anderen Jahrgangs zu führen.

Die Konsequenz aus diesen Überlegungen wäre, dass in Vorstudien eine dichtere Verlinkung der Jahrgänge über mehr gemeinsame Testlets erfolgen sollte. Erst wenn sichergestellt ist, dass diese Testlets keinen substanziellen Klassenstufen-DIF auf Aufgabenebene zeigen, kann man von unverzerrten Schätzungen der Leistungsdifferenz zwischen beiden Gruppen ausgehen und diese Aufgaben auch zur Abschätzung von längsschnittlichen Entwicklungstrends einsetzen. Methodisch kann hinterfragt werden, ob strenge Invarianztests (vgl. Steenkamp & Baumgartner, 1998) zur Feststellung nicht invarianter Itemschwierigkeiten und Itemtrennschärfen für einen Vergleich der beiden Klassenstufen notwendig sind. In den Ansätzen von Little, Slegers und Card (2006) sowie von De Jong, Steenkamp und Fox (2007) müssen Skalen nur durch eine Fixierung von Mittelwerten von Itemschwierigkeiten und

Itemladungen identifiziert werden. Variieren diese Parameter jedoch zwischen den Klassenstufen sehr stark, so ist nur eine eingeschränkte Interpretation der Skala möglich. Eine Abschätzung der Größe dieser Variabilitäten liefern aber gerade DIF-Varianzen (für Schwierigkeiten und Trennschärfen). In Anwendungen wie beispielsweise der Konstruktion eines Index für den Sozialstatus aus Besitzstandsvariablen sind jedoch DIF-Effekte für Vergleichsgruppen (Länder) bei bestimmten Items (z B. beim Besitz einer Klimaanlage in Ländern Nord- und Südeuropas) erwartbar und für Analysen in Rechnung zu stellen (vgl. May, 2006). Die Existenz von DIF schließt demzufolge die Vergleichbarkeit von Skalenwerten zwischen Subgruppen nicht aus. Sie birgt zwar die Gefahr einer erschwerten Interpretierbarkeit, gestattet aber andererseits, das interessierende Konstrukt in ganzer Breite zu erfassen.

Den in der Praxis vermutlich schwer realisierbaren Optimalfall der Verlinkung leistungsheterogener Gruppen stellt der Einsatz gemeinsamer, vollständig identischer Testhefte dar, wobei in verschiedenen Testheftversionen alle enthaltenen Testlets über alle denkbaren Positionen rotiert werden, um eine Konfundierung mit Positionseffekten ausschließen zu können. Diese Testhefte sollten änderungssensitive Items und Aufgaben beinhalten, für die ein Klassenstufen-DIF sowohl auf Item- als auch auf Aufgabenebene aufgrund von Pilotierungsergebnissen (oder ähnlichen Vorstudien) ausgeschlossen werden kann. Die Verwendung eines einzigen Testheftes zur Erfassung der Unterschiede zwischen den Klassenstufen eliminiert zwar Kontext- und Positionseffekte, schränkt aber aufgrund der eingeschränkten Anzahl eingesetzter Aufgaben stark die Validität ein und ist daher nicht unbedingt eine empfehlenswerte Alternative.

In jedem Fall sollte der Fokus der Anstrengungen bei der Verlinkung leistungsheterogener Gruppen verschiedener Jahrgänge auf der Konstruktion des Designs liegen, nicht auf der späteren Adjustierung der Daten um derartige Effekte (vgl. dazu etwa die Adjustierung um Bookleteffekte in PISA 2000: OECD, 2002; Mazzeo & von Davier, 2008). Vielmehr sollte das Ziel sein, bereits während der Studienplanung designbedingte Effekte wie Itemkontexte, Positionseffekte und Bookleteffekte so weit als möglich einzugrenzen.

Die hier thematisierten Probleme betreffen aber nicht nur die Verbindung von verschiedenen Jahrgangsstufen, sondern auch die Skalierungen identischer Jahrgangsstufen verschiedener Schularten im Sekundarstufenbereich (Hauptschule, Realschule, Gymnasium) und nehmen somit einen breiten Raum innerhalb der empirischen Bildungsforschung ein.

IV.5 Literatur

Ackerman, T. A., Gierl, M. J. & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, *22 (3),* 37-53.

AERA, APA & NCME (1999) siehe American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (NCME) (1999)

American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (NCME) (1999). *Standards for educational and psychological testing*. Washington, DC: APA.

Artelt, C. & Schlagmüller, M. (2004). Der Umgang mit literarischen Texten als Teilkompetenz im Lesen? Dimensionsanalysen und Ländervergleiche. In U. Schiefele, C. Artelt, W. Schneider & P. Stanat (Hrsg.), *Struktur, Entwicklung und Förderung von Lesekompetenz. Vertiefende Analysen im Rahmen von PISA 2000* (S. 169–196). Wiesbaden: VS Verlag für Sozialwissenschaften.

Baumert, J., Stanat, P. & Demmrich, A. (2001). PISA 2000: Untersuchungsgegenstand, theoretische Grundlagen und Durchführung der Studie. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider et al. (Hrsg.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 15-68). Opladen: Leske und Budrich.

Blais, J.-G. & Laurier, M. D. (1995). The dimensionality of a placement test from several analytical perspectives. *Language Testing*, *12*, 72-97.

Bock, R. D., Gibbons, R. & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*, 261-280.

Borg, I. & Staufenbiel, T. (2007). *Lehrbuch Theorien und Methoden der Skalierung* (4. Auflage). Bern: Hans Huber.

Bos, W., Lankes, E.-M., Schwippert, K., Valtin, R., Voss, A., Badel, I. & Plaßmeier, N. (2003). Lesekompetenzen deutscher Grundschülerinnen und Grundschüler am Ende der vierten Jahrgangsstufe im internationalen Vergleich. In W. Bos, E.-M. Lankes, M. Prenzel, K. Schwippert, G. Walther & R. Valtin (Hrsg.), *Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich* (S. 69-142). Münster: Waxmann.

Bos, W., Valtin, R., Lankes, E.-M., Schwippert, K., Voss, A., Badel, I. & Plaßmeier, N. (2004). Lesekompetenzen am Ende der vierten Jahrgangsstufe in einigen Ländern der Bundesrepublik Deutschland im nationalen und internationalen Vergleich. In W. Bos, E.-M. Lankes, M. Prenzel, K. Schwippert, R. Valtin & G. Walther (Hrsg.), *IGLU. Einige Länder der Bundesrepublik Deutschland im nationalen und internationalen Vergleich* (S. 49-92). Münster: Waxmann.

Bos, W., Valtin, R., Voss, A., Hornberg, S. & Lankes, E.-M. (2007). Konzepte der Lesekompetenz in IGLU 2006. In W. Bos, S. Hornberg, K.-H. Arnold, G. Faust, L. Fried, E.-M. Lankes et al. (Hrsg.), *IGLU 2006. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 81-108). Münster: Waxmann.

Bradlow, E. T., Wainer, H. & Wang, X. (1999). A Bayesian Random. Effects Model for Testlets. *Psychometrika*, *64*, 153-168.

Bremerich-Vos, A. & Böhme, K. (2009). Lesekompetenzdiagnostik – die Entwicklung eines Kompetenzstufenmodells für den Bereich Lesen. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 219-249). Weinheim: Beltz.

Brennan, R. L. (2001). *Generalizability Theory*. New York: Springer.

Burnham, K. P. & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information – Theoretical Approach* (2nd Edition). New York: Springer.

Camilli, G., Wang, M. & Fesq, J. (1995). The effects of dimensionality on equating the law school admission test. *Journal of Educational Measurement*, *32*, 79-96.

Camilli, G. & Penfield, D. A. (1997). Variance estimation for differential test functioning based on Mantel-Haenszel statistics. *Journal of Educational Measurement*, *34*, 123-129.

De Jong, M. G., Steenkamp, J. B. E. M. & Fox, J.-P. (2007). Relaxing crossnational measurement invariance using a hierarchical IRT model. *Journal of Consumer Research*, *34*, 260-278.

DeMars, C. (2002). Incomplete data and item parameter estimates under JMLE and MML estimation. *Applied Measurement in Education*, *15*, 15-31.

DiBello, L., Roussos, L., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R Rao & S. Sinharay (Eds.), *Handbook of Statistics, 26*, (pp. 979-1030). Amsterdam: Elsevier.

Douglas, J. A., Roussos, L. A. & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement*, *33*, 465-484.

Finch, H. & Monahan, P. (2008). A bootstrap generalization of modified parallel analysis for IRT dimensionality assessment. *Applied Measurement in Education*, *21*, 119-140.

Gailberger, S. (2007). Lesen und Lesebewusstheit im Deutschunterricht. Versuch einer multidimensional basierten Leseprozesstheorie zur Förderung des Lesens im Deutschunterricht. In S. Gailberger & M. Krelle (Hrsg.), *Wissen und Kompetenz. Entwicklungslinien und Kontinuitäten in Deutschdidaktik und Deutschunterricht. Heiner Willenberg zum 65. Geburtstag gewidmet* (S. 149-173). Baltmannsweiler: Schneider Hohengehren.

Gorin, J.S. & Embretson, S.E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, *30*, 394–411.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22*, 144–149.

Hanson, B.A. (2002). *ICL Manual. IRT Command Language. Version 0.020301*. Zugriff am 25.10.2011 unter http://www.b-a-h.com/software/irt/icl/icl_manual.pdf

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139–164.

Jang, E.E. & Roussos, L. (2007). An investigation into the dimensionality of TOEFL using conditional covariance-based nonparametric approach. *Journal of Educational Measurement*, *44*, 1-21.

Jannaronne, R.J. (1997). Models for locally dependent responses: Conjunctive item response theory. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 465-480). New York: Springer.

Jude, N., Klieme, E., Eichler, W., Lehmann, R. H., Nold, G., Schröder, K., Thomé, G. & Willenberg, H. (2008). Strukturen sprachlicher Kompetenzen. In DESI-Konsortium (Hrsg.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* (S. 191-201). Weinheim: Beltz.

Kamata, A. & Bauer, D.J. (2008). A note on the relation between factor analytic and item response theory models. Structural Equation Modeling: *A Multidisciplinary Journal, 15*, 136-153.

Kim, H. R. (1994). *New techniques for the dimensionality assessment of standardized test data*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, Department of Statistics.

Kirisci, L., Hsu, T.-C. & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement*, *25*, 146-162.

Kolen, M. J. & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking. Methods and Practices* (2nd Edition). New York: Springer.

Köster, J. (2005). Wodurch wird ein Test schwierig? Ein Text für die Fachkonferenz. *Deutschunterricht*, *58 (5),* 34-39.

Little, T. D., Slegers, D. W. & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling*, *13*, 59–72.

Longford, N. T., Holland, P. W. & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 171–196). Hillsdale, NJ: Erlbaum.

Marsh, H. W., Balla, J. R. & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis. The effect of sample size. *Psychological Bullentin*, *103*, 391-410.

May, H. (2006). A multilevel Bayesian item response theory method for scaling socioeconomic status in international studies of education. *Journal of Educational and Behavioral Statistics*, *31*, 63-79.

Mazzeo, J., & von Davier, M. (2008). *Review of the Programme for International Student Assessment (PISA) test design: recommendations for fostering stability in assessment results.* Zugriff am 26.08.2011 unter http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=EDU/PISA/GB%282008%2928&docLanguage=En

McCoach, D. B. & Black, A. C. (2008). Evaluation of model fit. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel Modeling of Educational Data* (pp. 245-271). Charlotte, NC: Information Age Publishing.

McDonald, R.P. & Mok, M.C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research*, *30*, 23-40.

Monahan, P. O., McHorney, C. A., Stump, T. E. & Perkins, A. J. (2007). Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics*, *32*, 92-109.

Monseur, C. & Berezner, A. (2007). The computation of equating errors in international surveys in education. *Journal of Applied Measurement*, *8*, 323-335.

Muthén, L. K. & Muthén, B. O. (2006). *Mplus. User's guide. Fifth edition*. Los Angeles, CA: Muthén & Muthén.

OECD (2002). *PISA 2000. Technical Report*. Paris: OECD.

Penfield, R. D. & Algina, J. (2006). A generalized DIF effect variance estimator for measuring unsigned differential test functioning in mixed format tests. *Journal of Educational Measurement*, *43*, 295-312.

R Development Core Team (2008). *R: A language and environment for statistical computing*. Wien: R Foundation for Statistical Computing.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models. Applications and data analysis methods* (2nd Edition). Newbury Park, CA: Sage.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*, 207-230.

Reckase, M. D., Ackerman, T. A. & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement*, *25*, 193-203.

Robitzsch, A. (2009). Methodische Herausforderungen bei der Kalibrierung von Leistungstests. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik* (S. 42-107). Weinheim: Beltz.

Roussos, L. & Ozbek, O. (2006). Formulation of the DETECT Population Parameter and Evaluation of DETECT Estimator Bias. *Journal of Educational Measurement*, *43*, 215-243.

Roussos, L. & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*, 355-371.

Roussos, L. & Stout, W. (2004). Differential item functioning analysis: Detecting DIF item and testing DIF hypotheses. In D. Kaplan (Ed.): *The Sage handbook of quantitative methodology for the social sciences* (pp. 107-115). Thousand Oaks, CA: Sage.

Rupp, A. A. & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, *8*, 219-262.

Schiefele, U., Artelt, C., Schneider, W. & Stanat, P. (Hrsg.) (2004). *Struktur, Entwicklung und Förderung von Lesekompetenz. Vertiefende Analysen im Rahmen von PISA 2000*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2005). *Bildungsstandards im Fach Deutsch für den Primarbereich (Jahrgangsstufe 4) – Beschluss vom 15.10.2004*. München: Wolters Kluwer.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2006). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring – Beschluss vom 02.06.2006*. München: Wolters Kluwer.

Shealy, R. & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*, 159–194.

Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.

Spiegelhalter, D., Thomas, A., Best, N. & Lunn, D. (2003). *WinBUGS. Version 1.4. User Manual*. MRC Biostatistics Unit.

Spinner, K. H. (2004). Lesekompetenz in der Schule. In U. Schiefele, C. Artelt, W. Schneider & P. Stanat (Hrsg.), *Struktur, Entwicklung und Förderung von Lesekompetenz. Vertiefende Analysen im Rahmen von PISA 2000* (S. 125-138). Wiesbaden: VS Verlag für Sozialwissenschaften.

Steenkamp, J.-B. & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*, 78-90.

Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201-210.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*, 589–617.

Stout, W. (2002). Psychometrics: From practice to theory and back. *Psychometrika*, *67*, 485-518.

Stout, W., Froelich, A. G., Gao, F. (2001). Using resampling methods to produce an improved DIMTEST procedure. In A. Boomsma, M. A. J. van Duijn & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 357-375). New York: Springer.

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L. A. & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, *20*, 331-354.

Swanson, D. B., Clauser, B. E., Case, S. M., Nungester, R. J. & Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics*, *27*, 53-75.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement*, *27*, 159-202.

Tran, U. S. & Formann, A. K. (2009). Performance of parallel analysis in retrieving unidimensionality in the presence of binary data. *Educational and Psychological Measurement*, *69*, 50-61.

Wainer, H., Bradlow, E. T. & Wang, X. (2007). *Testlet response theory and its applications*. New York: Cambridge University Press.

Wainer, H., Sireci, S. G. & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement*, *28*, 197-219.

Walker, C. M., Azen, R. & Schmitt, T. (2006). Statistical versus substantive dimensionality: The effect of distributional differences on dimensionality assessment using DIMTEST. *Educational and Psychological Measurement*, *66*, 721-738.

Wang, X., Bradlow, E. T., Wainer, H. & Muller, E. S. (2007). A Bayesian method for studying DIF: A cautionary tale filled with surprises and delights. *Journal of Educational and Behavioral Statistics*, *33*, 363-384.

Winkelmann, H. & Böhme, K. (2009). Anlage und Durchführung der Pilotierung der Bildungsstandards. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik* (S. 31-41). Weinheim: Beltz.

Wirth, R. J. & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*, 58-79.

Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. A. (2007). *ACER ConQuest. Version 2.0*. Camberwell, Victoria: Australian Council for Educational Research Press.

Yen, W. M. & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 111-154). Westport: Praeger Publisher.

Zhang, J. (2007). Conditional covariance theory and detect for polytomous items. *Psychometrika*, *72*, 69-91.

Zhang, J. & Stout, W. (1999). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika*, *64*, 129-152.

3.5 Kompetenzbereich IV – Sprache und Sprachgebrauch untersuchen

Der Kompetenzbereich IV *Sprache und Sprachgebrauch untersuchen* nimmt sowohl in der Formulierung der Bildungsstandards durch die Kultusministerkonferenz als auch in seiner Operationalisierung unter Leitung des IQB eine gewisse Sonderrolle ein.

Diese ergibt sich dadurch, dass die Kompetenz Sprache und Sprachgebrauch zu untersuchen in den Bildungsstandards integrativ angelegt ist und eine enge Anbindung an die anderen Kompetenzbereiche fordert. Die Idee der Vernetzung wird in den Standards im Fach Deutsch für den Mittleren Schulabschluss ausdrücklich betont, so heißt es: „Der Bereich ‚Sprache und Sprachgebrauch' steht in Beziehung zu jedem der drei anderen Bereiche" (KMK, 2004, S. 7). Im entsprechenden Dokument der Kultusministerkonferenz für den Primarbereich wird die angestrebte Verknüpfung dadurch veranschaulicht, dass der Bereich *Sprache und Sprachgebrauch untersuchen* „quer" zu den Bereichen *Sprechen und Zuhören*, *Schreiben* sowie *Lesen – mit Texten und Medien umgehen* angelegt ist (vgl. KMK, 2005a, Abb. S. 7).

Zudem umfasst der Bereich *Sprache und Sprachgebrauch untersuchen* Teilkompetenzen, die sich hinsichtlich ihres Auflösungsgrades und ihrer Ausrichtung deutlich unterscheiden. So stehen sich Kompetenzaspekte gegenüber, die einerseits im Sinne von *Sprachhandlungskompetenz* einen stark kommunikationsbezogenen Charakter aufweisen und auf Inhalte und Funktion von Sprache abzielen, andererseits sind auch Aspekte der sprachsystematischen *Grammatikarbeit* vertreten, die als eher kleinschrittig und formal beschrieben werden können.

Die Bildungsstandards im Fach Deutsch für den Primarbereich weisen die Kompetenz Sprache und Sprachgebrauch zu untersuchen somit als zentralen und übergreifenden Kompetenzbereich aus, der die folgenden Subfacetten umfasst (vgl. KMK, 2005a, S. 13):

- sprachliche Verständigung untersuchen,
- an Wörtern, Sätzen, Texten arbeiten,
- Gemeinsamkeiten und Unterschiede von Sprachen entdecken sowie
- grundlegende sprachliche Strukturen und Begriffe kennen und verwenden.

In den Standards ist wiederholt von Sprachhandlungskompetenz die Rede. Es wird betont, dass das sinnvolle sprachliche Handeln der Schülerinnen und Schüler sowie der sorgfältige und angemessene Umgang mit Sprache von ausschlaggebender Bedeutung sind (vgl. KMK, 2005a, S. 8). Weiter heißt es in den Bildungsstandards: „In altersgemäßen, lebensnahen Sprach- und Kommunikationssituationen erfahren und untersuchen die Kinder die Sprache in ihren

Verwendungszusammenhängen und gehen dabei auf die inhaltliche Dimension und die Leistung von Wörtern, Sätzen und Texten ein“ (KMK, 2005a, S. 9). Es wird deutlich, dass der Fokus der Bildungsstandards auf dem kommunikativen und inhaltlichen Gehalt von Sprache und Sprachgebrauch liegt.

Gleichzeitig findet sich in den Bildungsstandards die Forderung, dass die Schülerinnen und Schüler „ […] über ein Grundwissen an grammatischen Strukturen, einen Grundbestand an Begriffen und Verfahren zum Untersuchen von Sprache“ verfügen sollen (KMK, 2005a, S. 9). Der Deutschunterricht soll den Kindern nicht nur erste Einsichten in Sprachstrukturen ermöglichen, sondern sie auch mit elementaren Fachbegriffen vertraut machen (KMK, 2005a, S. 6). Hierbei werden grammatisches Begriffswissen und formale Grammatikarbeit in den Standards allerdings nicht als Selbstzweck verstanden, sondern stellen ein Werkzeug dar, um die produktive und rezeptive Sprachhandlungsfähigkeit der Schülerinnen und Schüler zu entwickeln und ihre metasprachliche Aufmerksamkeit und Bewusstheit zu fördern.

Neben der Orientierung an der Sprachhandlungskompetenz der Kinder widmet sich der Deutschunterricht in der Primarstufe also auch der formalen Grammatikarbeit, die sich zumeist an einzelnen Wörtern und Sätzen orientiert und eher auf technische Aspekte wie die Bestimmung von Wortarten oder Satzgliedern abzielt. Aus fachdidaktischer Perspektive fassen Oomen-Welke und Kühn (2008) zusammen: „Demnach enthält der Kompetenzbereich ‚Sprache und Sprachgebrauch untersuchen‘ nicht nur den Bereich der Grammatik und der Grammatikarbeit, sondern er umfasst den Bereich des Wortschatzes und der Wortschatzarbeit mit der Wortbildung und Redensarten; er reicht bis auf die Textebene“ (Oomen-Welke & Kühn, 2008, S. 139). Obwohl also beide Zugänge in den Bildungsstandards angelegt sind, wird dem Anwendungsaspekt von Sprachgebrauch Vorrang gegenüber formalen Aspekten eingeräumt. Dieser Fokus wird in den Standards auch explizit formuliert, indem festgestellt wird: „Priorität hat das sinnvolle sprachliche Handeln der Schülerinnen und Schüler“ (vgl. KMK, 2005a, S. 8).

3.5.1 Das Kompetenzkonstrukt „Sprache und Sprachgebrauch untersuchen“ aus Sicht der Deutschdidaktik

Gleicht man die in den Bildungsstandards vorgenommene Konstruktbeschreibung mit der deutschsprachigen Fachliteratur ab, wird zunächst offenbar, dass in der sprachdidaktischen Diskussion eine große begriffliche Vielfalt bei gleichzeitiger Unschärfe in der Definition dieser Begrifflichkeiten herrscht. So werden im Zusammenhang mit dem hier betrachteten Kompetenzbereich von verschiedenen Autoren Begriffe wie „Sprachthematisierung“, „Sprachaufmerksamkeit“, „Sprachbewusstheit“, „metasprachliches Wissen“, „Sprachwissen“ und

Language Awareness verwendet (vgl. hierzu Oomen-Weke, 2003). Teilweise finden diese Begriffe parallel Anwendung, teilweise bemühen sich verschiedene Autoren durch die Wahl verschiedener Begriffe gewisse inhaltliche Unterschiede zu markieren. Worin diese Unterschiede jeweils konkret bestehen, bleibt mitunter unscharf. Ob daher die vorgenommenen Abgrenzungen in jedem Fall zwingend und für die Fachdiskussion erforderlich sind, bleibt offen.

In der fachdidaktischen Literatur allgemein üblich ist eine Gegenüberstellung von verschiedenen sprachbezogenen Wissensarten, die sich unterschiedlicher theoretischer Ursprünge und somit auch unterschiedlicher Begrifflichkeiten bedient (bspw. Anderson, 2001; Karmiloff-Smith, 1992). So wird *implizites* oder *prozedurales Wissen* (*knowing how*) von *explizitem* beziehungsweise *deklarativem Wissen* (*knowing that*) abgegrenzt (vgl. bspw. Behrens & Eichler, 2008; Bremerich-Vos & Böhme, 2009b; Oomen-Welke & Kühn, 2008). Auch im Rahmen der DESI-Studie wurden innerhalb des Bereichs Sprachbewusstheit einerseits das explizite (deklarative) Wissen über Sprache und andererseits die konkrete sprachliche Korrekturfähigkeit (prozedurales bzw. implizites Wissen) unterschieden (Eichler & Nold, 2006, S. 63). Ebenso ist diese Unterscheidung in früheren Phasen der Vergleichsarbeiten in der dritten Jahrgangsstufe (VERA-3) vorgenommen und für die Etablierung eines Kompetenzstufenmodells herangezogen worden (vgl. Isaac, Eichler & Hosenfeld, 2008). Die Gleichsetzung von implizitem oder prozeduralem Wissen mit sprachpraktischem Können einerseits sowie von explizitem oder deklarativem mit metasprachlichem Wissen andererseits erscheint plausibel, wird jedoch, beispielsweise durch die Unterscheidung verschiedener Formen expliziten Wissens, wie zum Beispiel durch Karmiloff-Smith (1992), mitunter in Frage gestellt (vgl. auch Andresen & Funke, 2003). Eine Bezugnahme auf die verschiedenen Aspekte sprachlichen Wissens findet in den Bildungsstandards nicht ausdrücklich statt. Entsprechend der oben dargestellten Unterscheidung von eher handlungsorientierten und eher formalen Kompetenzaspekten liegt der Fokus allerdings eher auf den prozeduralen Anteilen.

In der sprachdidaktischen Literatur wird im Zusammenhang mit dem hier diskutierten Kompetenzbereich verschiedentlich die Bedeutung der metasprachlichen Reflexion betont. Hierfür wird von einigen Autoren der Begriff der Sprachbewusstheit gebraucht: „Als Sprachbewusstheit wird die Bereitschaft und Fähigkeit bezeichnet, sich aus der mit dem Sprachgebrauch in der Regel verbundenen inhaltlichen Sichtweise zu lösen und die Aufmerksamkeit auf sprachliche Erscheinungen als solche zu richten“ (Andresen & Funke, 2003, S. 439). Sprachbewusstheit erfordere also eine Loslösung vom inhaltlichen Gehalt einer sprachlichen Mitteilung und sei insofern stets metasprachlich. Die Kinder müssten lernen, dass es möglich ist, eine Distanz zu ihrem eigenen und auch fremdem Sprachhandeln aufzubauen. Sprache und Sprachgebrauch seien dann nicht länger Bestandteil einer kommunikativen

Situation, sondern würden von ihr abgetrennt und zum Gegenstand einer bewussten Reflexion gemacht. Diese Loslösung vom Inhalt des Kommunizierten, um die Aufmerksamkeit auf die sprachliche Ebene zu lenken, falle insbesondere jüngeren Kindern oft schwer. Wesentlich für die Untersuchung von „spontaner" Sprachbewusstheit bei Kindern sei somit die Frage, inwiefern Thematisierungen sprachlicher Formen und Funktionen auch tatsächlich Sprache und nicht durch Sprache ausgedrückte Inhalte oder durch Sprache erzielte Wirkungen zum Gegenstand hätten (Andresen & Funke, 2003, S. 445).

Diese Sichtweise deckt sich meiner Ansicht nach nicht vollständig mit der in den Bildungsstandards beschriebenen Kompetenz. Auch bleibt für mich fraglich, wieso die Thematisierung von durch Sprache erzielter Wirkung kein Nachweis von Sprachbewusstheit sein sollte.

Die in den Bildungsstandards formulierte Forderung, dass sich die verschiedenen Kompetenzbereiche im Sinne eines integrativen Deutschunterrichts aufeinander beziehen sollten (vgl. KMK, 2005a, S. 8), trifft wie oben ausgeführt für den Bereich „Sprache und Sprachgebrauch untersuchen" in besonderer Weise zu. Eine wechselseitige Bezugnahme könnte sich in einer ausgeprägten Orientierung an Sprache als Text äußern (zur Klärung des Textbegriffs vgl. Abschnitt 3.4.1). Diese Textbezogenheit wird von Seiten der Sprachdidaktik auch explizit eingefordert. Gleichzeitig wird verschiedentlich beklagt, dass auch die Integration von Texten in die grammatische Unterrichtsarbeit beziehungsweise die Integration von Sprachbetrachtung in die unterrichtliche Untersuchung von Texten bislang ungenügend ist. So äußert beispielsweise Scherner (2003) zu diesem Thema: „ […], dass die grammatische Beschreibung der Sprache und die Texthaftigkeit bzw. Textförmigkeit sprachlicher Gebilde im unterrichtlichen wie im fachdidaktischen Denken wenig, wenn nicht gar nichts miteinander zu tun haben" (Scherner, 2003, S. 476).

Eine ausdrückliche Stellungnahme von Vertretern der Fachdidaktik zu den Bildungsstandards findet sich unter anderem bei Kühn (2008), der für den Kompetenzbereich „Sprache und Sprachgebrauch untersuchen" jahrgangsübergreifende Grundsatzpositionen postuliert, die wie folgt zusammengefasst werden können (vgl. Kühn, 2008, S. 197):

- Die Untersuchung von Sprache und Sprachgebrauch sollte nicht isoliert, sondern in Anbindung an die anderen Kompetenzbereiche erfolgen. Angestrebt wird eine Vernetzung der Kompetenzbereiche untereinander.
- Grammatikarbeit sollte stets das Ziel verfolgen, kommunikative Sprachhandlungskompetenz zu vermitteln. Kompetenzentwicklung bedeutet somit, das

mündliche oder schriftliche Sprachhandeln der Schülerinnen und Schüler oder die textrezeptiven oder textproduktiven Sprachkompetenzen zu fördern.

- Sprachhandlungskompetenz sollte durch die Arbeit an und mit authentischen Texten aufgebaut werden. Die Untersuchung von Sprache und Sprachgebrauch darf nicht auf die Wort- oder Satzebene reduziert werden.

Kühn schließt seine Ausführungen mit dem Fazit: „Der Kompetenzbereich ‚Sprache und Sprachgebrauch untersuchen' ist grundsätzlich integrativ, kommunikativ und textorientiert auszurichten" (Kühn, 2008, S. 197).

Die in den Bildungsstandards für den Bereich *Sprache und Sprachgebrauch untersuchen* formulierten Kompetenzaspekte decken sich in mancher Hinsicht jedoch nicht vollständig mit den in der Sprachdidaktik geäußerten Überlegungen. Zudem wird deutlich, dass es in der Bestimmung des Kompetenzkonstrukts hier wie dort einige Unschärfen gibt, die beispielsweise die Gewichtung formaler beziehungsweise struktursystematischer Grammatikarbeit oder den Stellenwert der Textorientierung betreffen. Eine explizite Unterscheidung verschiedener sprachbezogener Wissensarten, wie sie in der fachdidaktischen Literatur üblich ist, findet sich in den Bildungsstandards nicht.

Insgesamt scheint das zugrunde liegende Konstrukt weniger präzise charakterisiert zu sein als andere relevante Kompetenzkonstrukte, wie beispielsweise die Lesekompetenz, außerdem umfasst es eine große und auch heterogene Menge von Teilfähigkeiten. Eine Unschärfe in der Konstruktdefinition findet sich dabei sowohl in den Erläuterungen der Bildungsstandards als auch in der sprachdidaktischen Diskussion.

3.5.2 Überlegungen zur Konstruktoperationalisierung

Unter Berücksichtigung der angesprochenen Unschärfen und einiger Diskrepanzen zwischen der sprachdidaktischen Perspektive und den Formulierungen in den Bildungsstandards hinsichtlich des zugrunde liegenden Kompetenzmodells wurde bei der Testentwicklung im Bereich „Sprache und Sprachgebrauch untersuchen" die Einbeziehung einer möglichst breiten und vielfältigen Aufgabenpalette angestrebt. Dennoch wird mitunter eine Diskrepanz zwischen den in den Bildungsstandards formulierten Kompetenzen und den eingesetzten Testaufgaben beklagt: „Abgefragt wird überproportional häufig […], ob die Schüler(innen) grammatische Fehler korrigieren oder grammatische Formen ergänzen, ob sie grammatische Termini Sprachbeispielen zuordnen können oder umgekehrt Sprachbeispiele grammatischen Termini" (Gornik &

Granzow-Emden, 2008, S. 133). Kühn (2008) bezeichnet die eingesetzten Aufgaben als „rein kognitiv und reproduktiv ausgerichtet und nicht kompetenzorientiert" (S. 201). Auch wenn hier die inhaltliche Verwendung des Begriffs „kognitiv" von dem in der Psychologie und in der empirischen Bildungsforschung üblichen Gebrauch abweicht, so wird doch deutlich, dass Kühn die primäre Konzentration auf Begriffswissen ablehnt, da das Verstehen der Begriffsinhalte seiner Ansicht nach für die Aufgabenbearbeitung nicht erforderlich ist.

Die in der Sprachdidaktik an den aktuell eingesetzten Testaufgaben geäußerte Kritik bezieht sich also darauf, dass die Aufgaben nicht integrativ mit anderen Kompetenzbereichen vernetzt und ferner zu wenig kommunikativ und textorientiert seien.

Tatsächlich muss eingeräumt werden, dass die eingesetzten Testaufgaben die in den Standards benannten Kompetenzaspekte nur unvollständig abbilden (vgl. Bremerich-Vos & Böhme, 2009b). So fehlen beispielsweise Aufgaben, welche die Beurteilung der kommunikativen Angemessenheit sprachlicher Mittel thematisieren. Auch die in der fachwissenschaftlichen Diskussion wiederholt geforderte und in den Bildungsstandards selbst angelegte Fokussierung auf die sprachliche Betrachtung von *Texten* wurde in den Testaufgaben zur Überprüfung der Bildungsstandards bislang noch nicht in zufriedenstellendem Umfang umgesetzt. Dieses Problem beschränkt sich nicht nur auf den Primarbereich, sondern betrifft auch die Operationalisierung im Kompetenzbereich „Sprache und Sprachgebrauch untersuchen" in der Sekundarstufe I. Insgesamt ist also das Konzept der Sprachhandlungskompetenz in den Testaufgaben bislang nicht hinreichend repräsentiert.

Aufgrund der genannten Defizite in der Aufgabenkonstruktion weist auch das Kompetenzstufenmodell für den Bereich „Sprache und Sprachgebrauch untersuchen" (vgl. Bremerich-Vos & Böhme, 2009b) inhaltliche Lücken auf und beschränkt sich gezwungenermaßen auf jene Kompetenzbestandteile, die tatsächlich untersucht worden sind.

Es stellt sich die Frage, weshalb die Operationalisierung des Konstrukts hinter der Konstruktbeschreibung zurückbleibt. Hierfür sind aus meiner Sicht drei Gründe ausschlaggebend, die nachfolgend kurz ausgeführt werden.

1) Zunächst kann für den hier diskutierten Kompetenzbereich in geringerem Umfang auf Erfahrungen aus anderen National Educational Assessments (bspw. NAEP) oder internationalen Erhebungen (bspw. PISA, PIRLS) zurückgegriffen werden. Im Gegensatz zur Überprüfung der Lese- oder Schreibkompetenz ist die testdiagnostische Untersuchung der Kompetenz Sprache und Sprachgebrauch zu untersuchen weit weniger verbreitet und zumeist kein expliziter Gegenstand der Erhebung.

Eines der wenigen Beispiele, in denen (englische) Sprachkompetenz überprüft wird, ist die regelmäßig durchgeführte *ICAS*-Studie[1], welche vom *Educational Assessment Australia* (EAA) der *University of New South Wales* verantwortet wird. Die Domäne der Sprachkompetenz wird vom Schreiben und Rechtschreiben abgegrenzt und wie folgt charakterisiert: ICAS-English „assess reading and language skills in a range of contexts. The following aspects of texts are assessed and reported: reading for meaning in literary texts, reading for meaning in factual texts, textual devices, syntax and vocabulary" (EAA, 2010). Die Kompetenz Sprache und Sprachgebrauch zu untersuchen wird hier also ausschließlich in Anknüpfung an die Überprüfung des Leseverstehens auf der Basis von Texten getestet. Dies entspricht den oben genannten Forderungen einer integrativen Sichtweise und könnte somit beispielgebend sein. Zunächst müsste allerdings untersucht werden, inwieweit eine empirische Abgrenzung des Sprachgebrauchs vom Lesen literarischer und informierender Texte bei einer solchen integrativen Testung möglich wäre (vgl. Abschnitt 2.4 sowie unten).

Anders gestaltet sich die Kompetenzfeststellung in NAPLAN, dem australischen *National Assessment Program Literacy and Numeracy*, bei dem unter anderem die Domäne *Language Conventions* getestet wird, welche die Überprüfung von Rechtschreibung, Grammatik und Zeichensetzung umfasst (vgl. NAPLAN, 2010). Beispielaufgaben lassen vermuten, dass sich die Untersuchung von Sprache und Sprachgebrauch auf den Aspekt der sprachlichen Richtigkeit beschränkt; allerdings sind die verfügbaren Informationen hinsichtlich der Definition der getesteten Konstrukte beschränkt.

Bemerkenswert ist, dass verschiedene nationale Large-Scale-Programme zum Monitoring von Bildungssystemen die Wortschatzarbeit (*vocabulary*) als Aspekt der Sprachhandlungskompetenz explizit berücksichtigen, obwohl Sprachgebrauch an sich nicht untersucht wird. Der Wortschatz wird dann entweder der Untersuchung des Leseverstehens zugeordnet oder als eigenständiger Bereich betrachtet. So heißt es im *Reading Framework for the 2009 National Assessment of Educational Progress*: „Recognizing a growing body of research that supports the argument that vocabulary is crucial to reading comprehension, the 2009 NAEP Reading Assessment will include a measure of vocabulary" (National Assessment Governing Board, 2008, S. 65ff.).

[1] Das Akronym ICAS steht für *International Competitions and Assessments for Schools* und bezeichnet jährlich durchgeführte Vergleichsarbeiten in den Jahrgängen 3 bis 12, an denen Schulen in Australien, Neuseeland sowie der Pazifikregion gegen Bezahlung freiwillig teilnehmen können. Testgegenstand sind jeweils Mathematik, Naturwissenschaften, Englisch, Schreiben, Rechtschreibung sowie Computerkenntnisse. (http://www.eaa.unsw.edu.au/).

Die am IQB verfügbaren Testaufgaben für die Überprüfung des Kompetenzstandes im Bereich „Sprache und Sprachgebrauch untersuchen“ berühren den Bereich der Wortschatzarbeit nur sehr vereinzelt. Dies begründet sich dadurch, dass sich im entsprechenden Dokument der Kultusministerkonferenz keine Erwähnung des Begriffs „Wortschatz“ findet. Mitunter wird die produktive und rezeptive Wortschatzarbeit jedoch von Seiten der deutschsprachigen Fachdidaktik als integraler Bestandteil des Kompetenzbereichs „Sprache und Sprachgebrauch untersuchen“ betrachtet (vgl. Oomen-Welke & Kühn, 2008), häufig bleibt dieser Aspekt aber gänzlich unerwähnt.

In diesem Kontext sollte allerdings berücksichtigt werden, dass es sich in englischsprachigen Nationen oftmals um Wortschatzarbeit in der Zweitsprache (L2), nicht um Wortschatzarbeit in der Muttersprache (L1) handelt. Dennoch haben die Erweiterung des Wortschatzes und die Überprüfung seiner Angemessenheit und seines Umfangs in anderen nationalen Programmen einen deutlich höheren Stellenwert (vgl. Read, 2000).

2) Bei der Überführung eines Kompetenzkonstrukts in Testaufgaben können nach Messick (1989a, 1989b, 1994) verschiedene Fehler auftreten (vgl. Abschnitt III.4). Zum einen können wesentliche Aspekte des theoretisch definierten Konstrukts bei der Operationalisierung im Test nicht repräsentiert sein, diese Operationalisierung wäre dann unvollständig (construct-underrepresentation). Zum anderen kann der Test Aspekte messen, die nicht in der theoretischen Konstruktdefinition enthalten sind und daher konstruktirrelevante Varianz erzeugen (construct-irrelevant variance). Eine Abwägung zwischen diesen beiden möglichen Fehlerarten kann aus diagnostischer Sicht bedeuten, dass darauf verzichtet wird, den Kompetenzstand im Bereich „Sprache und Sprachgebrauch untersuchen“ vorrangig auf der Basis von Texten zu überprüfen, da dies immer eine Konstruktkonfundierung mit der Lesekompetenz zur Folge hätte. Vereinfacht ausgedrückt wird ein wesentlicher Bestandteil des Konstrukts in der Testung vernachlässigt, um sicherstellen zu können, dass die verbleibenden Konstruktbestandteile (annähernd) unverfälscht erfasst werden können. Dass ein Großteil der Items lediglich auf der Wort- beziehungsweise Satzebene angesiedelt ist und nicht auf der Arbeit an und mit einem Text basiert, hat somit teilweise – allerdings nicht ausschließlich – auch diagnostische Gründe.

3) Ein weiterer möglicher Grund bezieht sich auf den Umstand, dass bei der Entwicklung der Testaufgaben stets auch der Aspekt der Unterrichtsvalidität maßgeblich war – selbst wenn dies ausdrücklich nicht der Grundidee der Bildungsstandards entspricht. Bremerich-Vos und Böhme (2009b) räumen ein, dass derzeit der Schwerpunkt der Testung im Bereich „Sprache und Sprachgebrauch untersuchen“ auf dem Substandard „grundlegende sprachliche Strukturen und

Begriffe kennen und verwenden“ liegt, geben aber gleichzeitig zu bedenken, dass dieser Schwerpunkt dem entspricht, was weite Teile der Unterrichtsarbeit ausmacht: „Damit dürften wesentliche Aspekte dessen bezeichnet sein, was in diesem Bereich nach wie vor faktisch im Zentrum des Unterrichts steht“ (Bremerich-Vos & Böhme, 2009b, S. 379).

Da die Testaufgaben von Lehrkräften entwickelt werden, spiegeln sie immer auch die aktuelle Unterrichtspraxis beziehungsweise die auf Seiten der Lehrkräfte vorherrschende Vorstellung der jeweiligen Kompetenz wieder. Da – wie oben dargestellt – neben dem Aspekt der Sprachhandlungskompetenz in den Bildungsstandards auch die formale Grammatikarbeit verankert ist, sollte es nicht überraschen, dass Aufgaben zu formalen Kompetenzbestandteilen vorgeschlagen werden, zumal diese in der Konstruktion weniger Aufwand bedeuten dürften. Dass Lehrkräfte mit einer solchen Sichtweise nicht zwingend allein sind, verdeutlicht der Umstand, dass auch Vertreter der Fachdidaktik ausdrücklich fordern, explizites beziehungsweise deklaratives sprachliches Wissen, welches auf eine formale und sprachsystematische Grammatikarbeit abzielt, nicht zu vernachlässigen (vgl. bspw. Dürscheid, 2007; Gornik & Granzow-Emden, 2008).

Anscheinend vertritt auch Kühn (2008) die Ansicht, die zu Testzwecken konstruierten Aufgaben seien repräsentativ für den aktuell prototypischen Grammatikunterricht. So heißt es in einem Beitrag, in dem Kühn Testaufgaben aus VERA-3 2006 sowie der DESI-Studie analysiert: „Der Aufgabentyp zeigt, dass in der Praxis des Grammatikunterrichts Aufgaben dominieren, die einseitig auf die Formseite der Sprache ausgerichtet sind. Semantische, funktionale oder pragmatische Aspekte bleiben weithin ausgeklammert“ (Kühn, 2008, S. 202).

Ob – und falls ja inwieweit – die hier eingesetzten Testformate tatsächlich das widerspiegeln, was der Praxis des deutschen Grammatikunterrichts entspricht, kann an dieser Stelle nicht entschieden werden. Deutlich wird aber, dass die aktuell eingesetzten Aufgaben vermutlich dem entsprechen, was die derzeitige Unterrichtsarbeit im Kompetenzbereich „Sprache und Sprachgebrauch untersuchen“ prägt. Insofern dürften die eingesetzten Aufgaben weitgehend unterrichtsvalide sein. Die Diskrepanz zwischen dem in den Bildungsstandards und von Seiten der Fachdidaktik formulierten Verständnis des Kompetenzbereichs „Sprache und Sprachgebrauch untersuchen“ besteht somit nicht nur zu den eingesetzten Testaufgaben, sondern auch zur aktuellen Unterrichtswirklichkeit. Nichtsdestotrotz bleibt die von Kühn (2008) vorgetragene Kritik an den Testaufgaben treffend: Die Aufgaben sollten stärker integrativ, kommunikativ und textorientiert ausgerichtet werden – diese Forderung trifft aber in gleichem Maß auf die Unterrichtsarbeit zu.

3.5.3 Fazit

Die Kompetenz Sprache und Sprachgebrauch zu untersuchen ist in den Bildungsstandards als integrative Sprachhandlungskompetenz angelegt. Eine solche Sichtweise wird von Vertretern der Sprachdidaktik unterstützt, dennoch besteht keine vollständige Identität zwischen der Sichtweise der Bildungsstandards und den Positionen in der fachdidaktischen Literatur. Aufgrund dieser leichten Differenzen finden sich einige Unschärfen in der Konstruktdefinition. Die testdiagnostische Umsetzung entspricht zwar der unterrichtlichen Praxis, wird den Vorgaben der Bildungsstandards und den fachdidaktischen Erwartungen aber nur bedingt gerecht und sollte zukünftig stärker integrativ, kommunikativ und textorientiert erfolgen.

Somit können unter anderem die folgenden Forschungsdesiderate abgeleitet werden:

- Zunächst sollte auf der Ebene der Konstruktdefinition eine theoretische Abklärung des Verhältnisses von kommunikationsrelevanter und auf Inhalte orientierter Sprachhandlungskompetenz einerseits und der Arbeit mit grammatischen Termini andererseits erfolgen.
- Anschließend sollte in die Entwicklung einer stärker integrativ orientierten und auf Textarbeit basierenden Batterie von Testaufgaben für die Kompetenzüberprüfung im Bereich „Sprache und Sprachgebrauch untersuchen“ investiert werden. Die dabei entstehenden Testaufgaben müssten gleichzeitig die Forderung nach Integration der verschiedenen Kompetenzbereiche – konkret Zuhören und Lesen – sowie eine stärkere Konzentration auf die Arbeit an und mit Texten erfüllen.
- Zusätzlich sollte erwogen werden, die Wortschatzarbeit explizit als Subfacette, welche die Sprachhandlungskompetenz maßgeblich bestimmt, im Kompetenzbereich „Sprache und Sprachgebrauch untersuchen“ zu berücksichtigen und in die Konstruktdefinition und Testentwicklung einzubeziehen.

4 Abschließende Diskussion

Da in den empirischen Beiträgen die jeweils themenspezifischen Diskussionspunkte bereits aufgegriffen und ausführlich erörtert wurden, soll sich die abschließende Diskussion nur kurz einer integrierenden Zusammenschau der zentralen Befunde sowie der wesentlichen Schlussfolgerungen der Einzelbeiträge widmen.

Im Anschluss möchte ich solche Diskussionsthemen aufgreifen, die im Kontext der länderübergreifenden Bildungsstandards sowie der zunehmenden Relevanz und Verbreitung groß angelegter standardisierter Schulleistungsuntersuchungen in einem umfassenderen Rahmen zum Tragen kommen. Für die vorliegende Arbeit sind diese Themen insofern bedeutsam, als sie ihren Kontext näher erschließen.

4.1 Zusammenschau der empirischen Beiträge

4.1.1 Zur Untersuchung des Hörverstehens mittels schwierigkeitsbestimmender Merkmale

Der erste empirische Beitrag mit dem Titel „Zur Abgrenzung des Hörverstehens gegenüber dem Leseverstehen mit Hilfe schwierigkeitsbestimmender Merkmale bei der Entwicklung von Testaufgaben“ (Böhme, Robitzsch & Busè, 2010) thematisiert die Operationalisierung des Hörverstehens im Rahmen von Large-Scale-Assessments und stellt Ergebnisse zur Analyse schwierigkeitsbestimmender Merkmale in bildungsstandardbasierten Aufgaben und Items zur Einschätzung der Kompetenzstände in den Bereichen Zuhören und Lesen vor. Hierbei werden zunächst die theoretischen Grundlagen der Bestimmung des Hörverstehenskonstrukts dargestellt und das Konzept schwierigkeitsbestimmender Aufgabenmerkmale eingeführt. Im empirischen Teil des Beitrags werden die den Analysen zugrunde liegenden Personen- und Itemstichproben beschrieben und ferner die gewählten Analyseschritte der Skalierung sowie der Regressions- und Kommunalitätenanalyse erläutert.

Ein zentraler Befund dieser Studie ist, dass sowohl für das Hör- als auch für das Leseverstehen allgemeine Aufgabenmerkmale, die sich auf das generelle Textverstehen beziehen, bei der Erklärung der Itemschwierigkeit einen wichtigen Stellenwert einnehmen. Um jedoch größere Teile der Varianz in der Itemschwierigkeit erklären zu können, müssen zusätzlich solche Merkmale berücksichtigt werden, die spezifische Eigenschaften des Hör- beziehungsweise Leseverstehens beschreiben. Diese Befunde sind erwartungskonform, da in theoretischen

Analysen der beiden rezeptiven Sprachkompetenzen betont wird, dass die kognitiven Informationsverarbeitungsprozesse weitreichende Ähnlichkeiten aufweisen, auf der Ebene der Rezeption jedoch die verschiedenen Inputstimuli (visuell vs. auditiv) berücksichtigt werden müssen (vgl. Kürschner & Schnotz, 2008). Die Betrachtung schwierigkeitsbestimmender Aufgabenmerkmale kann daher als ein Erfolg versprechender empirischer Zugang verstanden werden, um theoretische Überlegungen zur Unterscheidung des Hör- und Leseverstehens zu validieren.

Problematisch bei der Arbeit mit schwierigkeitsbestimmenden Aufgabenmerkmalen ist jedoch der Umstand, dass nicht alle Merkmale in gleicher Weise für die untersuchten Items und Texte sinnvoll und zutreffend sind. Die Ergebnisse zum Leseverstehen verdeutlichen beispielsweise, dass einige Merkmale nur textsortenspezifisch modelliert werden können, da sie nur innerhalb einer bestimmten Textsorte bedeutsam sind. Es scheint daher angezeigt, in weiterführenden Studien auch Interaktionseffekte zwischen Itemmerkmalen und der Textsorte zu berücksichtigen.

Wesentliche Limitationen des ersten Beitrags betreffen das methodische Vorgehen und hier zum einen die Einschätzung des Vorliegens der schwierigkeitsbestimmenden Merkmale in den Aufgaben sowie Items und zum anderen den Elaboriertheitsgrad der gewählten Analysemethoden. Die Einschätzungen der schwierigkeitsbestimmenden Merkmale wurden von Studierenden im Rahmen der Anfertigung ihrer Masterarbeiten vorgenommen. Obwohl alle Studierenden intensiv geschult und im Verlauf des Ratingprozesses von Experten beraten wurden, ist es dennoch fraglich, inwieweit die uns vorliegenden Einschätzungen mit denen von Fachexperten übereinstimmen würden. Dies betrifft die generelle Frage nach der Urteilsübereinstimmung von Ratern bei der Bewertung von objektiv nicht eindeutig entscheidbaren Sachverhalten. Dieses Thema ist auch für die Beurteilung von komplexen schriftsprachlichen Leistungen hoch relevant und erfährt daher im zweiten empirischen Beitrag (Böhme, Bremerich-Vos & Robitzsch, 2009), der sich mit der Kodierung von Schreibaufgaben auseinandersetzt, eine vertiefte Würdigung.

Die zweite methodische Beschränkung des ersten Beitrags betrifft den Einsatz wenig elaborierter Analysemethoden. Diese sind zwar in der Tat recht basal, für die Beantwortung der aufgeworfenen Fragen aber dennoch adäquat. Im Bemühen um eine Optimierung des analytischen Vorgehens wäre zunächst eine Übertragung des regressionsanalytischen Zugangs auf den Bereich der probabilistischen Testtheorie in Form des *Linearen Logistischen Testmodells* (LLTM) von Fischer (1973) angezeigt. Einen weiteren Ansatzpunkt bieten mehrdimensionale IRT-Modelle, die eine Berücksichtigung der theoretischen Überlegung gestatten würden, dass Items selten als Indikatoren einer einzigen, klar separierbaren kognitiven Fähigkeit verstanden werden

können, sondern für ihre Lösung vielmehr ein komplexes Zusammenspiel einer Vielzahl kognitiver Fähigkeiten erforderlich ist. Um eine entsprechende Modellierung der Daten und somit eine adäquate Abbildung der dimensionalen Kompetenzstruktur vornehmen zu können, stellen MIRT-Modelle, welche die Struktur kategorialer, beobachtbarer Variablen in multiplen, kontinuierlichen, latenten Dimensionen repräsentieren, einen denkbaren Zugang dar.

Da die Frage nach der Struktur der betrachteten sprachlichen Kompetenzen von zentraler Bedeutung für die vorliegende Arbeit ist, wird ihr in den empirischen Beiträgen III (Böhme & Bremerich-Vos, 2009) und IV (Böhme & Robitzsch, 2009a) näher nachgegangen, wobei methodisch anspruchsvollere Zugänge als im ersten empirischen Beitrag gewählt werden.

4.1.2 Zur Kodierung von Schülertexten im Kompetenzbereich Schreiben

Der zweite empirische Beitrag mit dem Titel „Aspekte der Kodierung von Schreibaufgaben“ (Böhme, Bremerich-Vos & Robitzsch, 2009) verfolgt das Ziel, fachdidaktische Überlegungen zum Konstrukt der Schreibkompetenz mit diagnostischen Fragen der Operationalisierung und Bewertung von Schreibprodukten zu verknüpfen. Neben der Untersuchung der Reliabilität der durch Rater vorgenommenen Einschätzungen der Schülertexte handelt es sich aus diagnostischer Perspektive auch um die Problematik der Dimensionalität des Konstrukts, die sich beispielsweise in verschiedenen Varianten der Kodierung niederschlagen kann. Somit verfolgt dieser Beitrag auch die Absicht, unterschiedliche Strategien bei der Beurteilung von Schreibaufgaben vorzustellen und hinsichtlich ihrer Eignung in Large-Scale-Assessments zu analysieren.

Um einen Vergleich der hier diskutierten analytischen und holistischen Kodierstrategien unter besonderer Berücksichtigung der Interraterreliabilität leisten zu können, wurden Befunde für eine narrative Schreibaufgabe vorgestellt. Im Hinblick auf die erreichten Beurteilerübereinstimmungen zeigt sich für dichotome Variablen der analytischen Kodierstrategie ein insgesamt befriedigendes Befundmuster. Jedoch ergeben sich verschiedentlich Probleme aufgrund stark ungleicher Kategorienbesetzungen. Mehrstufige analytische Variablen, die zumeist dem allgemein sprachlichen Bereich entstammen, können mehrheitlich nicht mit ausreichender Reliabilität gemessen werden. Dieser Befund lässt sich in *Multi-Trait-Multi-Method-Modellierungen* unter Rückgriff auf Strukturgleichungsmodelle replizieren. Die für vier- und sechsstufige holistische Variablen ermittelten Intraklassenkorrelationen, die als mittlere Korrelation zweier Rater interpretiert werden können, betragen .67 bis .75, was auch im Vergleich mit der internationalen Literatur als Erfolg gewertet werden kann.

Beim Vergleich der holistischen und analytischen Kodierstrategie bei der Beurteilung des Inhalts eines narrativen Textes zeigt sich ein hoher korrelativer Zusammenhang, der je nach

gewähltem Raterdesign im Bereich von $r = .80$ bis .90 liegt. Eine analytische Summenvariable für die Bewertung des Inhalts misst im hier diskutierten Beispiel geringfügig reliabler als eine holistische Einschätzung des Inhalts.

Betrachtet man die Zusammenhänge der (hinreichend reliabel messbaren) analytischen Sprachvariablen untereinander, ergeben sich durchgängig moderate positive Korrelationen. Den höchsten Zusammenhang mit dem holistischen Gesamteindruck zeigt die Variable Wortschatz, sie ist auch in einer linearen Regression der beste Prädiktor. Alle Variablen des allgemeinen Sprachbereichs der analytischen Kodierung weisen deutlich geringere Halo-Effekte als holistische Einschätzungen der Schülertexte auf.

Befunde zur Dimensionalität des Konstrukts der Schreibkompetenz lassen einen engen Zusammenhang zwischen stilistischen und inhaltlichen Aspekten erkennen. Die Komponente der sprachlichen Richtigkeit (Rechtschreibung und Grammatik) nimmt jedoch eine Sonderstellung ein und lässt sich nicht ohne Weiteres als Bestandteil eines eindimensionalen Konstrukts fassen. Dies steht im Einklang mit Befunden für den Sekundarbereich im deutschsprachigen Raum (vgl. Neumann, 2007).

Hinsichtlich des Vergleichs der analytischen mit der holistischen Kodierstrategie lässt sich festhalten, dass beide Strategien potenziell geeignet sind, um zu validen Aussagen hinsichtlich der Schreibkompetenz zu gelangen. In Bezug auf das jeweils erfasste Schreibkonstrukt kann festgehalten werden, dass es erkenntnislogisch unplausibel ist, anzunehmen, mit beiden Strategien würden sich unterschiedliche Konstrukte messen lassen. Vielmehr stellen die beiden Strategien unterschiedliche methodische Zugänge zur Messung desselben Personenmerkmals dar.

Bezug nehmend auf die Befunde zur Interraterreliabilität und Beurteilerübereinstimmung kann man pointiert festhalten, dass man mitunter zwar genau misst, wie im Falle der dichotomen analytischen Variablen, aber nicht immer zwischen guten und schlechten Schülern unterscheiden kann. Letzteres ist jedoch das eigentliche Ziel. Im Allgemeinen fallen Maße, wie die prozentuale Übereinstimmung oder Cohens κ, umso schlechter aus, je mehr Stufen die verwendete (Rating-)Skala umfasst. Analytische Variablen erzielen somit naturgemäß höhere Übereinstimmungen in Maßen wie der prozentualen Übereinstimmung oder Cohens κ, gleichzeitig sind sie aber in stärkerem Umfang dem Problem ungleicher Kategorienbesetzungen unterworfen und erfassen darüber hinaus meist weniger Varianz und damit auch weniger wahre Varianz in den Schülerantworten. Primäres Interesse gilt aber dem Signal-Rausch-Verhältnis in den eingesetzten Variablen und somit der Frage nach dem Verhältnis von wahrer Varianz und Messfehlern. Diese kann nicht allein durch eine weitgehend unreflektierte Bestimmung der prozentualen Übereinstimmung oder von Cohens κ beantwortet werden.

Stellt man dieses Problem zunächst zurück, kann aber festgehalten werden, dass analytische Kriterien nur geringe Halo-Effekte und kleine Interkorrelationen aufweisen. Dies bedeutet, dass eine Differenzierung von Teilaspekten der Schreibkompetenz im Fall der analytischen Kodierung sehr gut gelingt.

Die holistische Kodierung hingegen zeigt in Varianzkomponentenzerlegungen sehr erfreuliche Ergebnisse im Hinblick auf die ermittelte wahre Varianz. Gleichzeitig schneidet sie aufgrund der Vielstufigkeit der eingesetzten Skalen in Bezug auf Maße wie die prozentuale Übereinstimmung oder Cohens κ deutlich schlechter ab. Möglicherweise sind hier aber Maße der Interraterreliabilität wie die Intraklassenkorrelation – für die zufriedenstellende Ergebnisse gefunden wurden – das angemessenere Kritierium. Problematisch bleibt im Fall der holistischen Kodierstrategie allerdings das hohe Ausmaß an Interkorrelation der Teilbereiche, für die eine differenzielle Erfassung intendiert war. Ferner sind die starken Halo-Effekte ein Problem, das in der künftigen Forschung und Anwendung stärker berücksichtigt werden muss.

Die Frage, welcher Kodierstrategie in zukünftigen Studien aufgrund der vorliegenden Befunde der Vorzug zu geben ist, kann hier nicht abschließend beantwortet werden. Vielmehr wird man je nach Anlass und Zielstellung der Studie abwägen müssen, mit welcher Strategie man den im Vordergrund stehenden Fragen beziehungsweise Intentionen adäquater gerecht werden kann. Ein analytischer Kodierzugang besitzt aus fachdidaktischer Sicht zahlreiche wertvolle Vorteile, die sich insbesondere auf eine detaillierte Erschließung des Schreibproduktes beziehen, durch die sehr konkrete Aussagen über den Kompetenzstand eines Kindes getroffen werden können. In kleineren, fachdidaktisch orientierten empirischen Untersuchungen zur Schreibkompetenz wird man daher dieser Strategie den Vorzug geben. Auch bei Fragen der Individualdiagnostik wird ein differenziertes, auf Teilkomponenten des Schreibproduktes fokussierendes Vorgehen bei der Bewertung die Methode der Wahl sein. Nur auf diese Weise können eine differenzierte Erfassung von Stärken und Entwicklungspotentialen erfolgen und Maßnahmen für eine gezielte Förderung eingeleitet werden. Es darf allerdings nicht übersehen werden, dass vor einem Einsatz der analytischen Kodierstrategie dem Problem begegnet werden muss, dass für die wichtigen analytischen Sprachvariablen derzeit noch keine befriedigende Reliabilität erreicht werden konnte.

Für die Hand von Lehrkräften, beispielsweise im Rahmen von Vergleichsarbeiten oder für die tagtägliche Korrekturarbeit, scheint die hier vorgestellte analytische Kodierung weniger geeignet zu sein. Zu umfangreich und auch fachlich zu komplex sind die einzelnen Kriterien gefasst. Zudem zeigen die Übereinstimmungsmaße, dass der hohe Zeitaufwand nicht zwingend mit einer höheren Messgenauigkeit einhergeht, sondern Unsicherheiten bei der Kodierung bestehen bleiben. Für den schulischen Kontext oder auch für die Korrektur von Schreibaufgaben

im Rahmen von Vergleichsarbeiten bieten sich daher möglicherweise holistische Variablen für Teilbereiche der Schreibkompetenz wie die hier vorgestellten vierstufigen Variablen zu Inhalt, Stil und sprachlicher Richtigkeit an. Diese holistischen Bewertungen sollten aber keinesfalls als Ad-hoc-Eindrücke verstanden, sondern im Sinne wohldefinierter und kriterial verankerter Bewertungsmaßstäbe eingesetzt werden. Diese Art der Bewertung von Schreibprodukten könnte eine sinnvolle Balance zwischen notwendigen Informationen für eine fundierte Leistungsbeurteilung und gezielte Förderung einerseits und dem durch Lehrkräfte leistbaren Arbeitsaufwand andererseits darstellen. Zudem kann die Bereitstellung von Benchmarktexten und Beschreibungen der jeweiligen Stufen die Professionalität der Lehrkräfte im Bereich der Diagnostik von Schreibaufgaben befördern.

Für die Entwicklung von Kompetenzstufenmodellen sowie für den Einsatz in großen Schulleistungsstudien, wie beispielsweise in den Ländervergleichen auf Basis der Bildungsstandards, bietet sich vermutlich der Einsatz eines holistischen Gesamteindrucks neben einer geringen Anzahl analytischer und holistischer Variablen für Teilbereiche der Schreibkompetenz an. Ähnlich wie der hier vorgestellte sechsstufige Gesamteindruck sollte eine solche holistische Skala auf allen Stufen durch klare Kriterien charakterisiert sein und durch Schülerbeispiellösungen illustriert werden. Auf diese Weise ließen sich alle Aufgaben einer Textsorte – und möglicherweise sogar textsortenübergreifende Aufgaben – in einem gemeinsamen Stufenmodell der Schreibfähigkeit abbilden. Nur durch einen einheitlichen, gemeinsamen Bewertungsmaßstab wäre außerdem die angestrebte Dokumentation von Entwicklungstrends leistbar.

Der mögliche Vorwurf, dabei bleibe unklar, welche Leistung auf Schülerseite eigentlich gemessen werde, kann anhand der im Beitrag präsentierten Befunde entkräftet werden. Man kann recht genau bestimmen, woraus sich der Gesamteindruck speist: Es sind dies die vierstufigen holistischen Variablen Inhalt und Stil. Betrachtet man den Beitrag der allgemeinen analytischen Sprachvariablen, so rückt hier in erster Linie der Wortschatz als Aspekt des Stils in den Blick.

Insgesamt sollten die hier ausgesprochenen Empfehlungen eines deutlich kommunizieren: Was für den Einsatz in Large-Scale-Assessments adäquat erscheint, muss keinesfalls auch im Schulalltag der Weg der Wahl sein. Umso wichtiger ist es daher, in den flächendeckenden Vergleichsarbeiten und in der Fortbildung der Lehrkräfte ein Spektrum Erfolg versprechender diagnostischer Ansätze anzubieten.

4.1.3 Zur Diagnostik der Rechtschreibkompetenz

Der dritte empirische Beitrag mit dem Titel „Diagnostik der Rechtschreibkompetenz in der Grundschule – Konstruktprüfung mittels Fehler- und Dimensionsanalysen“ von Böhme und Bremerich-Vos (2009) befasst sich mit der Evaluierung der Bildungsstandards im Bereich der Rechtschreibkompetenz. Hierfür wählten wir einen diagnostischen Ansatz, der sowohl quantitative Aussagen über das globale Kompetenzniveau als auch Aussagen auf der Ebene von qualitativen Fehleranalysen innerhalb der orthografischen Stufe des Rechtschreiberwerbs gestattet. Die Fehleranalysen basieren auf neun orthografietheoretisch plausiblen Fehlerkategorien, die in Anlehnung an die Aachener Förderdiagnostische Rechtschreibfehler-Analyse AFRA (vgl. Herné & Naumann, 2005) entwickelt wurden.

Empirisch zeigt sich, dass eine Diagnostik der globalen Rechtschreibfähigkeit unter Berücksichtigung der Anzahl richtig geschriebener Wörter zum Ende der dritten und vierten Jahrgangsstufe erfolgreich geleistet werden kann. Eindimensionale Rasch-Skalierungen auf Wortebene liefern für alle vier entwickelten Testformen im Hinblick auf die Modellpassung gute Ergebnisse. Im Vergleich der Jahrgangsstufen drei und vier zeigt sich eine mittlere Leistungsdifferenz von $d = 0.70$, wobei zum Ende der vierten Jahrgangsstufe leichte Deckeneffekte auftreten. Erwartungsgemäß weisen Mädchen in beiden Jahrgangsstufen einen Leistungsvorsprung gegenüber den Jungen auf. Im Mittel ergibt sich hier ein Effekt von $d = 0.20$.

Auf deskriptiver Ebene fällt die überwiegende Mehrzahl der Fehler in die neun definierten Fehlerkategorien, womit unterstrichen wird, dass die Diagnostik innerhalb der orthografischen Stufe dem tatsächlichen Erwerbsstand entspricht.

Die Befundlage lässt somit erkennen, dass es gelungen ist, Testinstrumente zu konstruieren, die unter den Bedingungen des Large-Scale-Assessments eine reliable Messung der globalen Rechtschreibfähigkeit von Schülerinnen und Schülern der dritten und vierten Jahrgangsstufe gestatten. Diese Testverfahren beziehen sich auf Rechtschreibphänomene, die für die orthografische Phase des Erwerbs relevant sind.

Darüber hinaus galt unser Interesse der Frage, inwieweit sich für ein orthografietheoretisch fundiertes neundimensionales Modell der Rechtschreibkompetenz empirische Belege finden lassen. Hier zeigten sich für höherkomplexe Modellvarianten keine befriedigenden Ergebnisse. Eine Prüfung der Mehrdimensionalität auf Fehlerebene gemäß den neun orthografietheoretisch definierten Fehlerkategorien mittels einer Hauptkomponentenanalyse der tetrachorischen Korrelationen, mehrdimensionaler Rasch-Skalierungen sowie mithilfe kognitiver Diagnosemodelle ergibt jedoch keine Evidenz für neun Dimensionen. Somit zeigen verschiedene Analysemethoden übereinstimmend, dass eine

hinreichend reliable Messung jeweils homogener und voneinander separierbarer Dimensionen auf der Basis der verfügbaren Daten nicht möglich ist. Theoriegeleitete und explorative Analysen legen eine eindimensionale Lösung nahe. Dieser Befund stimmt mit Ergebnissen der aktuellen Literatur überein, in der beispielsweise für das Niederländische – das mit dem Deutschen sehr nahe verwandt ist – für den Primarbereich eine eindimensionale Struktur der Rechtschreibkompetenz nachgewiesen wurde (vgl. Notenboom & Reitsma, 2003).

Das Ergebnis der Eindimensionalität der Rechtschreibkompetenz kann verschiedene Gründe haben. So könnten Mängel bei der Operationalisierung von eigentlich zutreffend konzeptualisierten Konstruktdimensionen vermutet werden. Diese Mängel könnten sich in einer insgesamt und auch pro Testform zu geringen Zahl an verwendeten Indikatoren für einzelne Dimensionen niederschlagen. Ferner zeigt sich, dass die Varianzen der verschiedenen postulierten Dimensionen stark voneinander abweichen und oftmals sehr gering sind. Dieses Problem wirkt sich auch auf die mehrdimensionalen Skalierungen aus. Durch diese messmethodischen Erschwernisse können aber nicht die über alle Dimensionen und Testversionen hinweg problematischen Ergebnisse erklärt werden.

Ein weiterer möglicher Ansatzpunkt ist die Überlegung, dass mit der Fokussierung auf regelgeleitete Schreibungen eine Unterschätzung des Stellenwertes des orthografischen Speichers einhergeht. Es liegt nahe zu vermuten, dass die Schülerinnen und Schüler gegen Ende der Grundschulzeit in einem höheren Maße als angenommen Wörter als Ganzheiten aus dem lexikalischen Speicher abrufen. In diesem Fall würden die Lupenstellen eines Wortes dadurch überlagert, dass das Wort als Ganzes erinnert wird (wenn auch nicht notwendigerweise in der konventionellen Schreibung). Die Fehler in einem Wort wären dann nicht lokal unabhängig voneinander.

Dieser Vermutung sind Böhme und Robitzsch (2009b) nachgegangen, indem sie Testleteffekte in die IRT-Modelle aufgenommen haben, wodurch jedes Wort als ein Testlet behandelt wird. Mithilfe der Q3-Statistik nach Yen (1984, 1993), die Residualkorrelationen berücksichtigt und somit ein Maß für die lokale Abhängigkeit von Items liefert, konnten Böhme und Robitzsch (2009b) aber lediglich moderate Effekte der Einbettung der Fehlerlupenstellen in Wörter nachweisen. Somit kann auch dieser Ansatz nicht als generelle Erklärung herangezogen werden.

In weiterführenden Analysen zeigten Böhme und Robitzsch (2009b) mittels einer Varianzkomponentenzerlegung, die neben der wahren Schülerfähigkeit auch Interaktionen zwischen der Schülerfähigkeit und der Wortschwierigkeit sowie zwischen der Schülerfähigkeit und der Schwierigkeit der Lupenstellen berücksichtigt, dass der Anteil für die Interaktion zwischen Schüler und Wort deutlich größer ist als der für die Interaktion zwischen Schüler und

Lupenstelle. Dies kann so interpretiert werden, dass der Stellenwert des Wortlernens den Stellenwert des „Regellernens“ für Fehlerlupenstellen, also Rechtschreibphänomene, überwiegt.

Berücksichtigt man jedoch in Komponentenmodellen (*Uni-* bzw. *Multidimensional Componential IRT Models*, Hoskens & DeBoeck, 1995, 1997, 2001) die lokale Abhängigkeit von Lupenstellen innerhalb von Wörtern im Sinne von möglicherweise auftretenden Interaktionen innerhalb von Wörtern, so zeichnet sich eine mehrdimensionale Struktur ab.

Die Antwort auf die Frage nach der Dimensionalität des Konstrukts der Rechtschreibkompetenz lässt sich somit nicht eindeutig beantworten, scheint aber von der Elaboriertheit und Komplexität der Modellierung nicht unabhängig zu sein. Möglicherweise könnte sich an dieser Stelle der Vorwurf auftun, dass hier die Komplexität der Modellierung so lange gesteigert wurde, bis die ursprünglich intendierten Ergebnisse gefunden wurden. Dieser Überlegung ist allgemein zuzustimmen: Es kann nicht der Weg der Wahl sein, die Elaboriertheit und Komplexität der Modellierung so lange zu erhöhen, bis sich gewünschte Befunde ergeben. Im hier vorliegenden Fall sind die gewählten komplexen Komponentenmodelle jedoch theoretisch plausibel und daher angemessen.

Die hier berichteten Arbeiten sind auch praktisch relevant, da ein Kompetenzstufenmodell für den Bereich der Orthografie, das sich allein auf die Anzahl richtig geschriebener Wörter sowie die empirischen Schwierigkeiten dieser Wörter beschränkt, ausschließlich exemplarische Aussagen zu den Kompetenzständen von Schülerinnen und Schülern machen kann, die nicht verallgemeinerbar und damit wenig informativ sind. Angestrebt werden daher Aussagen hinsichtlich bestimmter Rechtschreibphänomene und die Beantwortung der Frage, auf welcher Kompetenzstufe eines Modells sich welche Phänomene verorten lassen. Um die Entwicklung eines solchen Modells zu leisten, ist es erforderlich, die hier berichteten Forschungsarbeiten weiter voranzutreiben und abzuklären, inwieweit die dimensionale Struktur der Rechtschreibkompetenz in Abhängigkeit von der gewählten Modellierung berücksichtigt werden kann.

4.1.4 Zu methodischen Aspekten der Erfassung der Lesekompetenz

Bei der Entwicklung des von Bremerich-Vos und Böhme (2009a) vorgestellten Kompetenzstufenmodells für den Bereich Lesen wurden in unserer Forschergruppe zahlreiche psychometrische Besonderheiten diskutiert, die im Rahmen aktueller messmethodischer Debatten zur Kompetenzdiagnostik und zur Etablierung von Kompetenzstufenmodellen interessante und relevante Fragen und Probleme betreffen. Daher verfolgt der vierte empirische Beitrag mit dem Titel „Methodische Aspekte der Erfassung der Lesekompetenz“ (Böhme &

Robitzsch, 2009a) das Ziel, aktuelle methodische Herausforderungen zu beleuchten und einige verallgemeinerbare Schlussfolgerungen für das Large-Scale-Assessment abzuleiten. Dies betrifft beispielsweise die Fragen nach methodischen Ansätzen zur Bestimmung der Konstruktdimensionalität (Abschnitt IV.2) oder Aspekte des differenziellen Itemfunktionierens in leistungsheterogenen Gruppen (Abschnitt IV.3).

Die in diesem Beitrag durchgeführten Dimensionsanalysen auf der Basis parametrischer und nichtparametrischer Verfahren können dahingehend interpretiert werden, dass zwar keine klar eindimensionale Struktur identifiziert wurde, aber auch keine eindeutig mehrdimensionale Struktur zu erkennen ist, die eine theoretisch plausible Differenzierung verschiedener Subfacetten gestatten würde.

Dies legt nahe, für die Beschreibung der Kompetenzstände der Schülerinnen und Schüler ein einheitliches Kompetenzstufenmodell zu wählen, wobei auf den einzelnen Stufen Hinweise auf Besonderheiten der Verstehensleistungen für einzelne Textsorten eine sinnvolle Ergänzung darstellen können. Die Etablierung separater Stufenmodelle für verschiedene Subdimensionen der Lesekompetenz wird durch die empirische Befundlage eher nicht gestützt. Entsprechend der hier vorgestellten Befunde wurde für die bildungsstandardbasierte Beschreibung der unterschiedlichen Kompetenzstände von Schülerinnen und Schülern am Ende des Primarbereichs im Kompetenzbereich Lesen ein Stufenmodell entwickelt, das die Lesekompetenz als homogenes Konstrukt auffasst, innerhalb der Stufenbeschreibungen aber auf Besonderheiten von literarischen und Sachtexten sowie ferner auf kontinuierliche und diskontinuierliche Stimuli Bezug nimmt (vgl. Bremerich-Vos & Böhme, 2009a).

Im Hinblick auf das differenzielle Itemfunktionieren zwischen dritter und vierter Jahrgangsstufe veranschaulichen die präsentierten Befunde, dass das Ausmaß des Klassenstufen-DIFs zwischen den beiden Jahrgängen insgesamt eher moderat ausfällt. Zerlegt man die Effekte in einen Testlet- und einen Itemanteil, so ist der deutlich größere Effekt auf das Testlet, also spezifische Aufgaben beziehungsweise Stimuli, zurückzuführen. Möglicherweise führen diskontinuierliche Sachtexte in Form komplexer Tabellen generell zu einer Benachteiligung von Drittklässlern.

Bei der Untersuchung des differentiellen Funktionierens von Items sollte generell berücksichtigt werden, dass bei der Erfassung der Lesekompetenz Items in Aufgaben geschachtelt sind und DIF sowohl auf Item- als auch auf Aufgabenebene zutage treten kann. Zeigt eine bestimmte Aufgabe auf Ebene des Testlets eine spezifische Bevorzugung oder Benachteiligung, so erweisen sich möglicherweise nicht nur ein oder zwei Items für die Verlinkung als unbrauchbar, sondern die gesamte Aufgabe. Das Maß einer hinreichenden Verlinkung sollte daher nicht nur die absolute Anzahl der Verlinkungsitems sein, sondern auch

die Anzahl der Verlinkungsaufgaben, da Verlinkungsfehler aus den beiden Varianzkomponenten der Aufgaben- und Itemebene entstehen.

Die Konsequenz aus diesen Überlegungen wäre, dass bei der Messung der Lesekompetenz in Vorstudien eine dichtere Verlinkung von leistungsheterogenen Gruppen über mehr gemeinsame Testlets erfolgen sollte, als in der Hauptstudie benötigt werden. Erst wenn sichergestellt ist, dass die eingesetzten Testlets keinen substanziellen DIF auf Aufgabenebene für die interessierenden Gruppen zeigen, kann man von unverzerrten Schätzungen der Leistungsdifferenz zwischen beiden Gruppen ausgehen und diese Aufgaben auch zur Abschätzung von längsschnittlichen Entwicklungstrends einsetzen.

In Bezug auf die Verlinkung der untersuchten Jahrgangsstufen drei und vier mit gemeinsamen Items ergibt sich, dass eine hinreichend große überlappende Itemmenge unbedingte Voraussetzung für die valide Bestimmung des Leistungsunterschieds zwischen den leistungsheterogenen Gruppen ist. Die Wahl der konkreten Analysemethode zur Verlinkung (getrennte Skalierung mit anschließendem Equating vs. gemeinsame Skalierung mit Hintergrundmodell) scheint eher sekundär. Von entscheidender Bedeutung ist, dass im Testdesign eine hinreichend dichte Verlinkung der betreffenden Gruppen durch gemeinsam bearbeitete Items sichergestellt wird.

Den in der Praxis vermutlich schwer realisierbaren Optimalfall der Verlinkung leistungsheterogener Gruppen stellt der Einsatz gemeinsamer, vollständig identischer Testhefte dar, wobei in verschiedenen Testheftversionen alle enthaltenen Testlets über alle denkbaren Positionen rotiert werden sollten, um eine Konfundierung mit Positionseffekten ausschließen zu können. Diese Testhefte sollten änderungssensitive Items und Aufgaben beinhalten, für die ein DIF-Effekt sowohl auf Item- als auch auf Aufgabenebene aufgrund von Pilotierungsergebnissen ausgeschlossen werden kann.

In jedem Fall sollte der Fokus der Anstrengungen bei der Verlinkung leistungsheterogener Gruppen stets auf der Konstruktion des Designs und nicht auf der späteren Adjustierung der Daten liegen. Vielmehr sollte das Ziel sein, bereits während der Studienplanung designbedingte Effekte wie Itemkontexte, Positionseffekte und Bookleteffekte so weit wie möglich einzugrenzen.

Die Befunde zur Verlinkung von leistungsheterogenen Gruppen – ebenso wie die soeben dargestellten Überlegungen – beziehen sich im vorliegenden Fall auf die Messung der Lesekompetenz in den Jahrgangsstufen drei und vier. Die hierbei thematisierten Probleme betreffen aber nicht nur die Verbindung von verschiedenen Jahrgangsstufen, sondern beispielsweise auch die Skalierungen identischer Jahrgangsstufen verschiedener Schularten im Sekundarstufenbereich (Hauptschule, Realschule, Gymnasium). Auch die Clusterung von Items

in Aufgaben, die im Kompetenzbereich Lesen durch die Verwendung von Lesetexten als gemeinsame Grundlage für alle Items einer Aufgabe besonders prominent ist, trifft in identischer Weise für die Verwendung auditiver Textstimuli bei der Diagnostik im Kompetenzbereich Zuhören zu. Die hier berichteten Befunde können somit in verschiedener Hinsicht übertragen werden und besitzen somit hohe Relevanz für die empirische Bildungsforschung.

4.2 Einschätzung der Bildungsstandards für das Fach Deutsch im Primarbereich

4.2.1 Die Umsetzung der Bildungsstandards für das Fach Deutsch in Testaufgaben

Eine wesentliche Eigenschaft der länderübergreifenden Bildungsstandards, die bereits im Rahmen der so genannten Klieme-Expertise (Klieme et al., 2007, S. 9) thematisiert wurde, ist ihre Konkretisierbarkeit durch Aufgaben. Dass hiermit neben einer Nutzbarmachung der Bildungsstandards für Lernaufgaben und damit für die Unterrichtspraxis natürlich auch die Umsetzung in Testaufgaben zum Zweck der Leistungsprüfung gemeint ist, wird ebenfalls bereits in der Klieme-Expertise angesprochen (Klieme et al, 2007, S. 81f.). Die Überführbarkeit der Bildungsstandards in standardisierte Testverfahren soll ein Mittel der schulübergreifenden Qualitätssicherung und -entwicklung darstellen.

Die Bildungsstandards für das Fach Deutsch werden der Forderung nach einer prinzipiellen Umsetzbarkeit in Testaufgaben allerdings nur bedingt gerecht. Auch aus der Perspektive der Deutschdidaktik ist dieses Problem für das Fach Deutsch besonders prominent (vgl. Spinner, 2008).

In jedem Kompetenzbereich finden sich Standards, die einer Operationalisierung im Rahmen einer standardisierten Leistungstestung – die bislang ausschließlich als Gruppentestung im Paper-Pencil-Format erfolgt – nicht zugänglich sind (vgl. hierzu Abschnitt 2.2.2). Exemplarisch wurden in Abschnitt 3.4.3 für den Bereich der Lesekompetenz solche Standards aufgeführt, die bei der Überprüfung der Erreichung der Bildungsstandards nicht in Testaufgaben umgesetzt werden konnten, ferner werden hierfür Gründe benannt.[1]

[1] Für die Teilbereiche Schreiben und Rechtschreibung findet sich eine Auseinandersetzung mit der Testbarkeit der zugehörigen Standards in den jeweiligen empirischen Beiträgen.

Wie bereits in Abschnitt 2.1.1 angedeutet, sollte die Operationalisierung sprachlicher Konstrukte aus meiner Sicht allerdings nicht als grundsätzlich unmöglich eingeschätzt werden. Die ausschlaggebende Frage ist vielmehr, auf welche Weise die derzeit begrenzten testmethodischen Möglichkeiten erweitert werden müssten, um auch die komplexen und vielschichtigen sprachlichen Kompetenzen in ihrer gesamten Breite erfassen zu können. Probleme bei der Überführung einzelner Standards beziehungsweise einzelner Kompetenzaspekte in Testaufgaben begründen sich

(1) zum einen in Limitationen der Testsituation,
(2) zum anderen durch Schwierigkeiten bei der intersubjektiv konstanten Bewertung von Schülerleistungen und
(3) oftmals in der Konzeption der Konstrukte beziehungsweise der Standards selbst.

Ad 1)
Durch die derzeitig vorherrschende Durchführung von Paper-Pencil-Tests in Gruppensitzungen mit begrenzter Testdauer verbietet sich der Einsatz verschiedener alternativer Leistungsmessungen, die aus diagnostischer Sicht den erforderlichen Gütemaßstäben genügen können. Dies betrifft insbesondere den Einsatz computergestützter Diagnostik, die sich auch auf kleinere Substichproben beschränken könnte.

Vorstellbar wäre hier unter anderem die Arbeit in Sprachlaborkabinen, welche eine Ausgabe und Aufzeichnung von Sprache einzelner Schülerinnen und Schüler ermöglichen würde. Auf diese Weise wäre die Simulation von sprachlicher Interaktion in dialogischen Situationen oder die Schaffung sinnvoll eingebetteter monologischer Sprechanlässe möglich. In einem solchen Setting könnten Standards aus dem Kompetenzbereich *Sprechen* und hier beispielsweise konkret die Standards „an der gesprochenen Standardsprache orientiert und artikuliert sprechen", „Sachverhalte beschreiben" oder „Begründungen und Erklärungen geben" umgesetzt werden (vgl. KMK, 2005a, S. 9f.).

Ebenfalls denkbar wäre der Einsatz von computergestützter Diagnostik im Bereich des freien Schreibens. Hierzu findet sich in dem Teilbereich „über Schreibfertigkeiten verfügen" ein eigener Standard: „den PC – wenn vorhanden – zum Schreiben verwenden und für Textgestaltung nutzen" (KMK, 2005a, S. 10). Zunächst müsste dennoch sichergestellt werden, dass Schülerinnen und Schüler der Primarstufe über genügend Computererfahrung verfügen, ihre Texte auch mit Hilfe einer Computertastatur verfassen zu können. Sollte diese Fragen jedoch positiv entschieden werden können, böte sich in einem solchen Testsetting die Möglichkeit, auch das Überarbeiten von Texten diagnostisch umzusetzen (vgl. Abschnitt IV.4), da dieses unproblematisch im Hintergrund protokolliert werden könnte, ohne dass es für die Kinder eine

Einschränkung darstellen würde. Nachfolgend sollte untersucht werden, inwieweit bei handschriftlichem und computerbasiertem Verfassen der Texte von einem einheitlichen Konstrukt auszugehen ist.

Ad 2)

Bei der Untersuchung sprachlicher Fähigkeiten ist es tatsächlich mitunter schwierig, eine Schülerleistung ohne Weiteres dichotom als richtig oder falsch zu bewerten (vgl. Kammler & Knapp, 2002; Spinner, 2008).

Der Wunsch, auch die (erfolglosen) Bemühungen der Kinder zu würdigen, ist insbesondere im Rahmen der Vergleichsarbeiten in der dritten Jahrgangsstufe sehr prominent. So heißt es zum Beispiel bei Bartnitzky: „Die Ergebnisse zeigen kein wirkliches Leistungsbild, im Gegenteil: Sie verkennen die Leistungen der Kinder und tun ihnen Unrecht“ (Bartnitzky, 2008, S. 17). In Bartnitzkys Argumentation findet sich der richtige Gedanke, dass Grundschulkinder mitunter Fehler machen, die belegen, dass sie bereits bestimmte Kompetenzaspekte erworben haben. So führt Bartnitzky (2008) unter anderem an, dass im Rahmen der Vergleichsarbeiten 2008 zu dem Nomen „Kampf“ ein Adjektiv gebildet werden musste und verschiedene Kinder Adjektivbildungen wie „kämpferlich“ oder „kämpfrig“ anboten. In der Tat ist dies ein Nachweis dafür, „dass die Kinder über einen Adjektivbegriff verfügen und adjektivgemäß Wörter bilden“ können (Bartnitzky, 2008, S. 16). Es steht zu hoffen, dass Lehrkräfte diese Leistung im schulischen Alltag identifizieren und gezielt kanalisieren können.

Im Rahmen einer standardisierten Leistungsüberprüfung ist zunächst ausschlaggebend, welche Kompetenz erfasst werden soll. Geht es darum, zu einem konkreten Nomen das zugehörige Adjektiv zu benennen? Dann handelt es sich primär um einen Wortschatztest und die oben angebotenen Antworten sind zwar kluge Fehler, aber dennoch Fehler und die Schülerantworten somit *eindeutig falsch*. Ist das Ziel der Diagnose jedoch die kompetenzorientierte Frage, ob die Kinder über das Konzept der Adjektivbildung verfügen, dann sind die gegebenen Antworten ein Ausweis der zu diagnostizierenden Kompetenz und sollten als *eindeutig richtig* bewertet werden. Unabhängig von dieser Differenzierung könnte erwogen werden, stärker auf *Partial-Credit-Modelle* zurückzugreifen und damit auch Lösungen zu berücksichtigen, die nur teilweise korrekt sind oder ein geringeres Maß an Verständnis als „optimale Lösungen“ vermuten lassen. Dass durch eine solche Klassifizierung allerdings mit zunehmender Anzahl verfügbarer Kategorien auch der Einfluss intersubjektiv inkonsistenter Urteile zunimmt, liegt auf der Hand.

Zusammenfassend besteht für das hier dargestellte Beispiel aus meiner Sicht kein Diskussionsbedarf hinsichtlich der Entscheidbarkeit der Klassifizierung der Schülerantworten. Vielmehr ist es erforderlich, klar zu definieren, welches Konstrukt erfasst werden soll. Ferner

sollte stärker zwischen der Leistungsbewertung im schulischen Alltag und in standardisierten Leistungstests differenziert und diese Differenz anerkannt werden.

Anders verhält es sich mit Schülerleistungen, bei denen die subjektive Komponente der Bewertung tatsächlich sehr prominent ist. Dies bezieht sich beispielsweise auf das freie Schreiben und hier unter anderem auf die Frage, was genau „verständliches" Schreiben ausmacht, da hierbei stets berücksichtigt werden muss, *was* für *wen* verständlich sein soll. Schreiben Schülerinnen und Schüler beispielsweise eine Geburtstagseinladung an ihre Freunde, so besteht in der Leistungsbewertung, die nicht durch die Freunde, sondern durch geschulte Kodiererinnen und Kodierer erfolgt, zu den subjektiven Vorlieben zusätzlich die Herausforderung einer Perspektivübernahme, die ebenfalls unterschiedlich gut gelingen dürfte. Dass die Entwicklung von Auswertungsschemata und die Schulung der Beurteiler große, aber lösbare Herausforderungen umfassen, wurde in Beitrag II thematisiert.

Auch hinsichtlich der intersubjektiven Beurteilung von erbrachten Schülerleistungen besteht keine prinzipielle Unmöglichkeit, sondern vielmehr das Erfordernis, Zeit und Kraft zu investieren, um befriedigende Ergebnisse zu erhalten.

Ad 3)

Viele der in den Bildungsstandards benannten Kompetenzen können als hoch komplexe und intersubjektiv nur schwer zu klassifizierende Konstrukte charakterisiert werden. Die Entwicklung von Bildungsstandards, Kompetenzmodellen und Testinstrumenten stellen beispielsweise dann eine schwierige Herausforderung dar, wenn es um „literarästhetische Erfahrungsqualitäten des Umgangs mit Literatur geht" (vgl. Spinner, 2008, S. 313). Aktuelle Bemühungen, das literarästhetische Empfinden als Konstrukt zu definieren und messbar zu machen (vgl. bspw. Frederking, 2008; Frederking et al., 2009), gelten innerhalb der Deutschdidaktik sowie der pädagogisch-psychologischen Diagnostik als (noch) nicht zufriedenstellend gelöst.

Das Problem liegt hier ebenso wie bei den entsprechenden Bildungsstandards ansatzweise in der Entwicklung geeigneter Testinstrumente und ebenso in Ansätzen bei der Bewertung der Schülerantworten. Wie soll man beispielsweise messen, ob jemand „lebendige Vorstellungen beim Lesen und Hören literarischer Texte entwickel[t]" (KMK, 2005a, S. 11)? Soll man diese schriftlich wiedergeben oder zeichnen lassen? Sollen die so fixierten Vorstellungen dann dahingehend bewertet werden, ob der Beurteiler die zugrunde liegenden literarischen Texte wiederkennt oder ob die schülerseitigen Vorstellungen seinen eigenen ähneln? Aus meiner Sicht liegt das Problem dieses und ähnlicher Standards weniger in seiner testdiagnostischen Umsetzung oder in der intersubjektiv einheitlichen Bewertung der generierten Schülerantworten, vielmehr

liegt das Problem in der Unschärfe des Konstrukts selbst. Was ist eine „lebendige Vorstellung“ und wie definiert sich in diesem Kontext „entwickeln“?

Die Lösung dieses Problems könnte zum einen darin liegen, auf der Ebene der Begriffsbildung und Konstruktdefinition klarer und präziser zu formulieren. Ein anderer Ansatz könnte darin bestehen zu akzeptieren, dass die Bildungsstandards im Fach Deutsch sowohl für den Primarbereich als auch für die Sekundarstufe einzelne Standards umfassen, die nicht in Tests überführbar sind. Die Überführung scheitert jedoch nicht primär an testdiagnostischen Limitationen, sondern an der Unschärfe der Standards selbst. Solche Standards sind zweifellos dazu geeignet, den Unterricht zu bereichern und einen wesentlichen Beitrag zur Bildung der Schülerinnen und Schüler zu leisten, für eine Einschätzung, ob die in ihnen formulierten Ziele erreicht wurden, eignen sie sich jedoch nicht. Dass dies einen Widerspruch zu den postulierten Zielen der Bildungsstandards darstellt, ist unbestritten.

4.2.2 Konsequenzen der Bildungsstandards und regelmäßiger Testdurchführungen für den Deutschunterricht

Innerhalb der Deutschdidaktik bestehen Strömungen, die sich gegen die Etablierung des Kompetenzbegriffs und die Manifestierung der Bildungsstandards für das Fach Deutsch richten (vgl. bspw. Steinbrenner, 2007). Hintergrund dieses Standpunktes sind neben Gefühlen der Fremdbestimmung vorrangig die Angst vor der Marginalisierung zentraler Gegenstände des Faches, insbesondere im Hinblick auf Literatur beziehungsweise Literaturdidaktik. In diesem Zusammenhang formuliert Spinner:

> Die Tatsache, dass eine angemessene Formulierung und Überprüfung von Bildungsstandards zum Literaturunterricht so schwierig ist, führt derzeit zur Gefahr, dass die spezifisch literarischen Erfahrungs- und Lernprozesse im Deutschunterricht marginalisiert […] werden. […] Es ist deshalb wichtig, das Bewusstsein für jene Bildungsaspekte, die sich der Formulierung als kompetenzorientierte Standards nicht fügen wollen, offen zu halten. Sonst gehen wichtige Aspekte von Bildung verloren. (Spinner, 2008, S. 322)

Dass die gegenwärtige Orientierung an kognitiven Kompetenzbestandteilen sowie am Literacy-Konzept bewirkt, dass die pragmatischen und funktionalen Aspekte sprachlicher Kompetenzen dominieren und in der aktuellen Debatte um die Leistungsfähigkeit schulischer Bildung primär von Interesse ist, inwieweit die Schülerinnen und Schüler in sprachlicher Hinsicht gesellschaftlich handlungsfähig sind, ist aus meiner Sicht eine zutreffende Einschätzung.

Im Rahmen der Bildungsstandards der Kultusministerkonferenz für das Fach Deutsch (KMK, 2004, 2005a, 2005b) wird Literatur beziehungsweise literarische Bildung in eine Vielzahl pragmatisch orientierter sprachlicher Kompetenzen eingereiht und damit, gemessen an ihrem im Deutschunterricht tradierten Stellenwert, in den Hintergrund gedrängt. Während dies für die Bildungsstandards im Primarbereich aufgrund der Zielstellung dieser Unterrichtsphase – der Erwerb schriftsprachlicher Kompetenzen – kaum ins Gewicht fällt, wird die Verschiebung der Prioritäten in der Sekundarstufe I offensichtlich und findet in den momentan in der Entwicklung befindlichen Bildungsstandards für die Sekundarstufe II ihren Höhepunkt. Der gymnasiale Deutschunterricht der Abiturphase war traditionell und ist auch heute in der überwiegenden Mehrzahl der Fälle *Literaturunterricht*. Dies ist aus meiner Sicht ein schützenswertes Gut, da die Auseinandersetzung mit Literatur in der Tat wesentliche Bestandteile einer umfassenden Bildung ausmacht und für die Persönlichkeitsentwicklung wichtige Chancen bereithält (vgl. Spinner, 203, 2008).

Kann es aber der Weg der Wahl sein, dem Literaturunterricht Vorrang gegenüber der Vermittlung einer pragmatisch und funktional ausgerichteten Sprachhandlungskompetenz einzuräumen? Wiegt die Schulung literarästhetischen Empfindens schwerer als die schulische Vermittlung von argumentativer Schreibkompetenz? Sind einzelne Vertreter der allgemeinen Erziehungswissenschaft sowie der Literaturdidaktik nicht vielleicht Vorstellungen von Bildung und Persönlichkeitsentfaltung verhaftet, die lediglich eine gymnasiale Minderheit der deutschen Schülerschaft in den Blick nehmen und mit den tatsächlichen (Leistungs-)Verhältnissen des Bildungssystems kaum vereinbar sind (vgl. hierzu Kämper-van den Boogaart, 2003)?

Bei der Bevorzugung des Literaturunterrichts gegenüber der Vermittlung sprachlicher Handlungskompetenz handelt es sich meines Erachtens um eine hochgradig elitäre Denkfigur, die den realen Bedürfnissen und Möglichkeiten einer Vielzahl – wenn nicht der Mehrzahl – der Schülerinnen und Schüler des deutschen Bildungssystems nicht gerecht wird. Fraglich ist für mich jedoch nicht *ob*, sondern lediglich *zu welchem Zeitpunkt und in welchem Umfang* ein auf literarische Bildung zielender Unterricht stattfinden sollte. Eine solche Differenzierung und Priorisierung von Bildungszielen darf allerdings nicht dazu führen, dass Literaturunterricht ausschließlich in leistungsstarken Klassen der gymnasialen Oberstufe angeboten wird. Vielmehr sollte aus meiner Sicht zentral sein, dass neben einer literarischen Bildung durch schulischen Deutschunterricht auch – und zwar *zuerst* – die für eine gleichberechtigte gesellschaftliche Teilhabe erforderliche sprachliche Handlungsfähigkeit aller Schülerinnen und Schüler hergestellt wird.

4.2.3 Fachspezifische vs. fächerübergreifende Kompetenzen

Ein anderes, aus meiner Sicht nach wie vor ungeklärtes Problem der Bildungsstandards für das Fach Deutsch ist der fächerübergreifende Stellenwert der verkehrs- beziehungsweise muttersprachlichen Kompetenzen. Die Bildungsstandards nehmen für sich in Anspruch, fächerspezifische Kernelemente zu thematisieren (vgl. hierzu Abschnitt 2.2.2; Klieme et al., 2007; Köller, 2008, 2010). Die Notwendigkeit sprachlicher Handlungsfähigkeit ist allerdings nicht spezifisch für das Fach Deutsch, sondern stellt für die Lernprozesse in nahezu sämtlichen Schulfächern sowohl in mündlicher als auch in schriftlicher Form eine elementare Voraussetzung dar. Folgerichtig heißt es in den Bildungsstandards für das Fach Deutsch im Primarbereich: „Die Beherrschung der deutschen Sprache ist für alle Kinder eine wichtige Grundlage für ihren Schulerfolg, denn Sprache ist in allen Fächern Medium des Lernens“ (KMK, 2005, S. 6). Insofern lassen sich die (verkehrs-)sprachlichen Kompetenzen in eine Gruppe von fächerübergreifenden Schlüsselkompetenzen einreihen, zu denen beispielsweise auch die Problemlösekompetenz, Lernkompetenz oder soziale Kompetenz gezählt werden (vgl. Klieme, Stanat & Artelt, 2001). Im Unterschied zu diesen nichtfachlichen Kompetenzen können sie aber als fächerübergreifende und gleichzeitig fachliche Kompetenzen charakterisiert werden (vgl. Böhnisch, 2008). Entscheidend ist jedoch, dass es sich um fächerübergreifend relevante Kompetenzen handelt. Inwieweit die Förderung der verkehrssprachlichen Kompetenzen nun aber allein Gegenstand des regulären Deutschunterrichts ist – obwohl sie auch für andere Fächer eine bedeutsame Lernvoraussetzung darstellen – wird in den Bildungsstandards nicht hinreichend thematisiert. Hier wäre es wünschenswert herauszustellen, dass die im Deutschunterricht auf- beziehungsweise auszubauende sprachliche Handlungsfähigkeit ein zentrales Ziel schulischer Bildung insgesamt darstellt. Aufgrund seiner Bedeutung als Lernvoraussetzung könnten für die Verwirklichung dieses Ziels zusätzliche – zeitliche und personelle – Ressourcen eingefordert werden. Eine (realistischere) Alternative wäre die Stärkung der fächerübergreifenden Zusammenarbeit, die seit Jahren – vergeblich – gefordert wird.

4.3 Nutzen und Nutzbarmachung von Kompetenzdiagnostik

4.3.1 Kompetenzdiagnostik als Werkzeug der Qualitätsentwicklung im Bildungswesen

Eine Kompetenzdiagnostik, die im Rahmen von großen internationalen Schulleistungsstudien, dem nationalen Bildungsmonitoring oder auch im Kontext jährlicher flächendeckender Vergleichsarbeiten durchgeführt wird, muss letztlich daran gemessen werden, ob sie positive Entwicklungsimpulse in das Bildungswesen hineinträgt. Gleiches gilt für die Orientierung an den länderübergreifenden Bildungsstandards und ihre Forderung nach kompetenzorientiertem Unterricht.

Welcher Art könnten diese Entwicklungsimpulse kompetenzorientierter Leistungsdiagnostik mittels standardisierter Testinstrumente sein (vgl. bspw. EMSE-Netzwerk, 2008)?

Testinstrumente, die auf den länderübergreifenden Bildungsstandards basieren, steigern zunächst den Bekanntheitsgrad der Standards und tragen dazu bei, eine insgesamt erhöhte Aufmerksamkeit gegenüber den Erträgen schulischer Bildung zu sichern.

Kompetenzdiagnostische Befunde aus Schulleistungstests können, insbesondere im Rahmen der jährlichen flächendeckenden Vergleichsarbeiten (VERA), eine an den Bildungsstandards ausgerichtete *kriteriale Orientierung* hinsichtlich der Leistungsstände der Schülerinnen und Schüler liefern und damit Lernergebnisse jenseits des sonst üblichen Klassenkontexts einordnen. In diesem Kontext zeichnet sich ab, dass eine Auseinandersetzung mit den für den Primarbereich im Fach Deutsch vorliegenden Kompetenzstufenmodellen vorrangig im Rahmen der Ergebnisrückmeldung von VERA-3 erfolgt und die Lehrkräfte erst auf diesem Weg die Bildungsstandards als kriteriale Bezugsnorm kennen lernen.

Ferner gestatten kompetenzdiagnostische Testbefunde, die in so genannte *faire Vergleiche* (vgl. Fiege, Reuther & Nachtigall, 2011) eingebettet sind, eine Gegenüberstellung der Leistungen der eigenen Klasse oder Schule mit Leistungen in sozial vergleichbaren Schulen und damit die *Orientierung an einer sozialen Norm*. Auch in diesem Fall wird der Bewertungshorizont über die soziale Norm des eigenen Klassenkontexts hinaus erweitert. Speziell für Lehrkräfte, die an so genannten Brennpunktschulen unterrichten, bieten faire Vergleiche realistische Maßstäbe an und zeigen auf, was möglicherweise auch mit Schülerinnen und Schülern aus sozial benachteiligten Schichten oder Kindern, die nicht muttersprachlich Deutsch sprechen in zentralen Kompetenzbereichen, wie beispielsweise dem Lesen, erreicht werden kann und sollte.

Leistungsergebnisse aus standardisierten Testinstrumenten können somit als Vergleichsmaßstab für die eigenen diagnostischen Urteile der Lehrkräfte dienen und damit die diagnostische Kompetenz des pädagogischen Personals fördern. Dass ein solcher Förderbedarf auch hinsichtlich der Beurteilung von Schülerleistungen im Fach Deutsch im Primarbereich besteht, wurde beispielsweise von Lorenz und Artelt (2009) untersucht.

Die in den Testinstrumenten eingesetzten Aufgabenformate sowie die im Rahmen von VERA zur Verfügung gestellten didaktischen Erläuterungen können zum einen Ideen für die Leistungsüberprüfung im Unterricht liefern, beispielweise hinsichtlich der Leistungsbewertung, zum anderen gestatten sie die unterrichtliche Vertiefung und Weiterarbeit. So wurde zum Beispiel im Rahmen der Vergleichsarbeiten der dritten Jahrgangsstufe im Jahr 2010 die orthografische Kompetenz der Schülerinnen und Schüler untersucht. Hierbei kamen Korrekturaufgaben zum Einsatz, die anhand qualitativer Fehleranalysen ausgewertet werden konnten. Dies brachte den Lehrkräften nicht nur einen im Unterricht bislang wenig verwendeten Aufgabentyp näher, sondern eröffnete auch die Chance zur Auseinandersetzung mit einer profilorientierten Diagnostik im Bereich der Orthografie, die aus didaktischer Sicht als sehr wertvoll eingeschätzt wird (vgl. Bremerich-Vos et al., 2010).

Entscheidend ist in diesem Kontext aber auch, dass Testdiagnostik *allein* nicht die hochgesteckten Ziele der Verbesserung und Stabilisierung der Qualität des deutschen Bildungssystems erbringen kann. Oelkers und Reusser schreiben in diesem Kontext:

> Die Erträge eines Schulsystems zu kennen, ist eine zwar hilfreiche, jedoch nicht hinreichende Bedingung zu einer Qualitätsverbesserung. So wächst unser Wissen über die Leistungsergebnisse unserer Bildungssysteme derzeit schneller als das Wissen darüber, wie diese Erkenntnisse für eine Verbesserung der Angebots- und Prozessqualität genutzt werden können. (Oelkers & Reusser, 2008, S. 494)

Ferner betonen die Autoren, dass der Unterricht „das zentrale Medium der Qualitätssicherung" (Oelkers & Reusser, 2008, S. 494) darstellt. Es ist daher wichtig, dass neben den Anstrengungen, sehr gute bildungsstandardbasierte Testinstrumente zu entwickeln, auch in die Implementierung und damit in eine standardorientierte und kompetenzbezogene Unterrichtsentwicklung investiert wird. Allerdings hat in diesem Kontext „die Frage, inwieweit eine kompetenzorientierte Sicht auf Schülerleistungen ein tragfähiges Modell für Unterrichtsentwicklung bildet, einen zentralen Stellenwert" (Peek, 2008, S. 168).

Ob sich die intendierten Wirkungen aber tatsächlich entfalten und die erhofften Entwicklungsprozesse angeregt wurden beziehungsweise werden, kann an dieser Stelle nicht

beantwortet werden. Mit Peek hat nach wie vor Folgendes Geltung: „Empirisch tragfähige Evaluationsbefunde, inwieweit dieses Modell einer standard- und kompetenzorientierten Unterrichtsentwicklung nachweisbare Effekte auf die Kompetenzsteigerung bei Schüler(inne)n in fachlichem Lernen hat, stehen noch aus“ (Peek, 2008, S. 168).

Ausschlaggebend bleibt, nicht nur die methodischen und fachdidaktischen Grundlagen und damit die Qualität der Kompetenzdiagnostik in Deutschland beständig zu optimieren, sondern auch ihre Wirksamkeit als Werkzeug der Qualitätsentwicklung im Bildungswesen fest im Blick zu behalten. Nur wenn es gelingt, die Akzeptanz von und das Verständnis für (quantitative) empirische Kompetenzdiagnostik bei den Lehrkräften auszubauen, kann dieses Instrument positive Entwicklungsimpulse ausstrahlen und die mit der Einführung der länderübergreifenden Bildungsstandards intendierten Verbesserungen in der schulischen Bildung bewirken. Es kann nicht genügen, Konstrukte zu definieren, diese zu operationalisieren und empirische Befunde in Tabellen und Diagrammen zu arrangieren. Bildungsforschung ohne verwertbare Implikation für die pädagogische Praxis ist zu Recht Gegenstand der Kritik von Seiten der Pädagogik sowie der Fachdidaktiken. Ebenso notwendig ist allerdings die Bereitschaft, quantitativ-empirisches Arbeiten als Erkenntnis stiftenden Zugang anzuerkennen und sich für eine Auseinandersetzung mit den auf diesem Wege gewonnen Ergebnissen zu öffnen und sie in das alltägliche pädagogische Handeln zu integrieren.

Wesentliche Voraussetzung hierfür ist unter anderem die stärkere Berücksichtigung entsprechender Inhalte in den Lehramtsstudiengängen sowie im Studium der Erziehungswissenschaft (vgl. hierzu bspw. Merkens, 2006; Sailer, 2007).

4.3.2 Erwartungsdivergenz: System- vs. Individualdiagnostik

Ein zentrales Problem der Operationalisierung von Bildungsstandards stellen aus meiner Sicht die divergierenden Erwartungen bezüglich der in den letzten Jahren eingesetzten Formen der Kompetenzdiagnostik dar. Als Instrument der Bildungspolitik sollen Bildungsstandards zeitgleich verschiedene Lücken schließen. So ist ein wesentlicher Grund für die Entwicklung und Verabschiedung der Standards in dem wiederholt erwartungswidrig schlechten Abschneiden deutscher Schülerinnen und Schüler in internationalen Schulleistungsstudien zu sehen (vgl. Abschnitt 2.2.1). Ausschlaggebend ist hierbei, dass nicht die Leistungen einzelner Schülerinnen und Schüler als problematisch eingestuft wurden, sondern dass die Leistungsfähigkeit und Qualität des gesamten deutschen Bildungssystems als defizitär eingeschätzt wurde.

Während aus Perspektive der empirischen Bildungsforschung im weiteren Sinne (vgl. Bildungsrat, 1974) in aller Regel Aussagen im Hinblick auf die – mitunter vergleichende –

Leistungsfähigkeit *eines Bildungssystems* angestrebt werden, steht aus Sicht der Erziehungswissenschaft, der Deutschdidaktik sowie der pädagogisch-psychologischen Diagnostik traditionsbedingt eher der gelingende Kompetenzerwerb *bei konkreten und damit stets einzelnen Schülerinnen oder Schülern* im Fokus des Interesses. Diese unterschiedlichen Perspektiven verschiedener Wissenschaftsdisziplinen spiegeln sich auch in den Standpunkten der involvierten Akteure. Während Bildungspolitiker die Sichtweise des Systemmonitorings einnehmen und damit das Bildungssystem als Ganzes oder doch zumindest ganze Schulen im Blick haben, besteht auf Seiten der Lehrkräfte ein Interesse an ihren jeweiligen Schülerinnen und Schülern als einzelnen Individuen. Diese Unterschiede im Fokus führen zu methodischen Missverständnissen und sind meines Erachtens ein wesentlicher – wenn nicht der entscheidende – Grund für die teils heftig geäußerte Kritik an der um sich greifenden „Testeritis" (vgl. bspw. Bartnitzky, 2008).

Neben diesen konfligierenden Perspektiven verschiedener Wissenschaftsdisziplinen und Akteure innerhalb des Bildungssystems trägt auch die Erwartungshaltung der Bildungspolitik zu dieser Problemsituation bei, die von Seiten der empirischen Bildungsforschung mitunter nicht deutlich genug korrigiert wird. Bereits im Rahmen der Klieme-Expertise (Klieme et al., 2007) wurde festgehalten, dass Bildungsstandards unter anderem dazu dienen, eine Grundlage für die Erfassung und Bewertung von Lernergebnissen zu liefern. Mit ihrer Hilfe lässt sich zum einen feststellen, „inwieweit das Bildungssystem seinen Auftrag erfüllt hat (Bildungsmonitoring)", zum anderen ist es möglich, den Schulen „eine Rückmeldung über die Ergebnisse ihrer Arbeit" (S. 10) zur Verfügung zu stellen (Schulevaluation). Neben den Diagnosezielen des Bildungsmonitorings und der Schulevaluation können die Bildungsstandards grundsätzlich auch Hinweise für eine individuelle Diagnostik und Förderung geben. „Allerdings legt die Expertise Wert darauf, dass Tests, die im Bildungsmonitoring und für die Schulevaluation eingesetzt werden, solche Individualdiagnostik aus methodischen Gründen meist nicht erlauben. Von einer Verwendung der Standards bzw. standard-bezogener Tests für Notengebung und Zertifizierung wird abgeraten" (Klieme et al., 2007, S. 10).

Der ausschlaggebende Punkt wurde also bereits bei der Entwicklung und Verabschiedung der Bildungsstandards festgehalten: Für individualdiagnostische Zwecke können die Bildungsstandards selbst durchaus eine sinnvolle Grundlage bilden. Ist dies das erklärte Ziel, müssen aber auch Testverfahren entwickelt werden, die auf Individualdiagnostik ausgelegt sind. Testverfahren, die dem Monitoring eines Bildungssystems dienen, können für einzelne, konkrete Schülerinnen und Schüler keine verlässlichen Aussagen treffen.

Mitunter erscheint mir in diesem Zusammenhang auch die Rolle der empirischen Bildungsforschung als problematisch: Gelegentlich wird von Vertretern des Faches der Eindruck erweckt, dass tatsächlich *alles* machbar sei, was von Seiten der Bildungspolitik nachgefragt und

gewünscht wird. Dies betrifft beispielsweise den Wunsch, kausale Zusammenhänge beschreiben zu können, um aufzuklären, welche Maßnahmen beziehungsweise Faktoren guten Unterricht ursächlich bedingen. Ein selbstkritischer Blick offenbart an dieser Stelle, dass die empirische Bildungsforschung in Deutschland in mancherlei Hinsicht recht optimistisch eingestellt ist, welche Aussagen mit hinreichender Sicherheit getätigt werden können. Möglicherweise ist diese Einstellung in einigen Belangen *zu* optimistisch. Dies soll aber weder bedeuten, die deutsche empirische Bildungsforschung sei defizitär, noch dass sie sich chronisch selbst überschätze. Was allerdings von der Disziplin akzeptiert werden muss, ist die Tatsache, dass die empirische Bildungsforschung in Deutschland im internationalen Vergleich erst seit kurzem floriert und in diesem Sinne über beschränktes Erfahrungswissen verfügt.

4.3.3 Rückmeldeverfahren: Die (unzureichende) Nutzbarmachung der Ergebnisse von Kompetenzdiagnostik

Bereits in der Klieme-Expertise heißt es: „Konkretisiert in Testverfahren, werden Standards im Rahmen des Bildungsmonitorings und der Evaluation von Schulen angewandt. Sie dienen der Feststellung und Bewertung von Lernergebnissen und haben somit Rückmeldefunktion, mit der sie zur outputorientierten Steuerung beitragen“ (Klieme et al., 2007, S. 47). In diesem Kontext äußern Oelkers und Reusser (2008), dass „die Rückverwandlung von Output in (verbesserten) Input“ allerdings mehr umfassen müsse als „die bloße Rückmeldung der Ergebnisse an die Akteure“ (S. 494). Im Folgenden bezeichnen die Autoren das Thema einer zielführenden und handlungsleitenden Rückmeldung als „eine – wenn nicht *die* zentrale Aufgabe und Herausforderung der Implementation“ (Hervorhebung im Original; Oelkers & Reusser, 2008, S. 494).

Ergebnisrückmeldungen aus Schulleistungstests können dazu beitragen, Mängel in der Unterrichtsgestaltung in Klassen und an Schulen zu identifizieren. In diesem Sinne können sie die Basis für Veränderungen bilden und zur Schul- und Unterrichtsentwicklung beitragen. Voraussetzung hierfür ist allerdings, dass die Lehrkräfte vor Ort die Ergebnisrückmeldungen verstehen, reflektieren und zielführend in Handlung umsetzen können (und wollen). Tatsächlich stellt aber die Frage, wie die Ergebnisse standardisierter und bildungsstandardorientierter Leistungsmessungen so aufbereitet und kommuniziert werden können, dass die Lehrkräfte vor Ort diese Informationen nutzbringend in ihre Unterrichtsgestaltung einfließen lassen können, ein weitgehend ungeklärtes Problem dar. Ferner sind häufig hohe und undifferenzierte Erwartungen an die Wirksamkeit und Innovationskraft von Ergebnisrückmeldungen gebunden, denen dieses Instrument allen Erfahrungen nach in der Regel nicht gerecht werden kann. So äußern Vertreter

des EMSE-Netzwerks in einem Positionspapier: „Die Erwartung, dass allein die Bereitstellung von externen Evaluationsdaten eine praxiswirksame Diagnose- und Reflexionsfunktion ausübe und gleichsam im Selbstlauf unterrichtsentwickelnde Konsequenzen nach sich ziehe, bestätigt sich bislang nicht“ (EMSE-Netzwerk, 2008, S. 2). Darüber hinaus wird in der einschlägigen Literatur kritisch diskutiert, dass häufig unklar ist, welchem Ziel die Rückmeldung konkret dienen soll (vgl. bspw. Hosenfeld & Groß Ophoff, 2007).

Diese Undifferenziertheit und Unklarheit in Zielen und Erwartungen äußert sich unter anderem darin, dass die verschiedensten Formen von Ergebnisrückmeldungen häufig nicht sinnvoll unterschieden werden: „Ferner werden Schulrückmeldungen unabhängig von der Funktion der jeweiligen Untersuchung und ihrem Design unter einer Kategorie subsumiert, als wäre es bedeutungslos, ob die Ergebnisse aus Querschnitt- oder Längsschnittstudien, aus Stichproben- oder Totalerhebungen stammen“ (Watermann, Stanat, Kunter, Klieme & Baumert, 2003, S. 92).

Entscheidend ist somit zunächst die Art der Schulleistungsuntersuchung, welche der Ergebnisrückmeldung zugrunde liegt, also die Frage, ob es sich beispielsweise um stichprobenbasierte, internationale oder nationale Schulleistungsstudien oder um jährlich flächendeckend durchgeführte Vergleichsarbeiten (VERA) handelt. Hinsichtlich erstgenannter Studien wird von Seiten der empirischen Bildungsforschung eindringlich davor gewarnt, an Ergebnisrückmeldungen im Kontext stichprobenbasierter Large-Scale-Assessments zu hohe Erwartungen zu knüpfen. So äußern beispielsweise Watermann und Kollegen zu den Schulrückmeldungen bei PISA-2000, es gäbe keine überzeugenden Argumente für die Annahme, „Schulrückmeldungen würden automatisch zu Innovations- und Optimierungsprozessen an Schulen führen“ (Watermann et al., 2003, S. 107). Die jährlichen, flächendeckenden Vergleichsarbeiten (VERA) in der dritten und achten Jahrgangsstufe dienen jedoch ausdrücklich dem Ziel der Schul- und Unterrichtsentwicklung, weshalb die Ergebnisrückmeldungen eine Schlüsselrolle einnehmen und angenommen wird, dass sie einen unmittelbaren und handlungsleitenden Nutzen erbringen (vgl. Koch, Groß Ophoff, Hosenfeld & Helmke, 2006). Hierbei obliegt es allerdings den Schulen und Lehrkräften vor Ort, aus den zur Verfügung gestellten Informationen Handlungsimplikationen abzuleiten und Entscheidungen hinsichtlich der jeweils erforderlichen Maßnahmen zu treffen.

Der Weg von der Information zur Innovation ist allerdings „weit und beschwerlich“ (Helmke & Hosenfeld, 2005). Wesentliche Determinanten einer zielführenden Rezeption von Ergebnisrückmeldungen sind die Akzeptanz, die inhaltlich zutreffende Interpretation der Informationen sowie die Ableitung von Implikationen für die Unterrichtsgestaltung. Diese drei

Aspekte der Akzeptanz, Verständlichkeit und Nutzung für die Unterrichtsarbeit sollen nachfolgend kurz thematisiert werden:

Schneewind (2007) untersuchte den Umgang Berliner Lehrkräfte mit Ergebnisrückmeldungen aus Schulleistungsstudien mittels einer Interview- und einer Fragebogenstudie. Hierbei stellte sie fest, dass insgesamt eine hohe Akzeptanz unter den Lehrkräften zu verzeichnen war. Ferner stellte sich heraus, dass die Qualität der den Ergebnisrückmeldungen zugrunde liegenden Tests für die Akzeptanz und die Nutzung der Informationen von entscheidender Bedeutung ist. Hierbei muss allerdings nicht zwangsläufig eine Übereinstimmung zwischen der von den Lehrpersonen wahrgenommenen, subjektiven Qualität und der aufgrund objektiver Kriterien feststellbaren Qualität der Testinstrumente vorliegen (vgl. Schneewind, 2007). Dennoch ist die Qualität der eingesetzten Tests sowie der didaktischen Begleitmaterialen ein wichtiger Einflussfaktor hinsichtlich der Akzeptanz von VERA auf Seiten der Lehrkräfte, weshalb durch das IQB und die fachdidaktischen Kooperationspartner stetig Bemühungen unternommen werden, die Güte der Testinstrumente auf hohem Niveau zu stabilisieren.

Hinsichtlich der Verständlichkeit von Ergebnisrückmeldungen aus Schulleistungsstudien kann als gesichert gelten, dass Lehrkräfte – in Abhängigkeit der von ihnen unterrichteten Fächer und Schulstufen – teils über sehr geringe statistische Vorbildung verfügen. Dies führt dazu, dass bereits basale Begriffe der quantitativ-empirischen Datenanalyse, wie zum Beispiel Mittelwert und Streuung, für eine große Anzahl der Lehrpersonen inhaltsleere Worthülsen darstellen (vgl. Schneewind, 2007). Dementsprechend wird auch das Konzept von mit Unsicherheit behafteten, statistischen Aussagen von den Lehrpersonen häufig nicht durchdrungen. Dieses Problem wird drastisch gesteigert, sobald es um methodische Grundlagen der probabilistischen Testtheorie und der hierauf beruhenden Entwicklung von Kompetenzstufenmodellen geht.

Dass ein Mindestmaß an Verständnis für empirische Ergebnisdarstellungen allerdings eine zwingende Voraussetzung für die Interpretation und Nutzbarmachung der Resultate aus Leistungstests ist, wurde von Seiten der Bildungsadministration erst in der jüngeren Vergangenheit ausreichend berücksichtigt (vgl. bspw. Emmrich, 2010). Auch aus wissenschaftlicher Perspektive wurde angemerkt, dass der Frage einer gelingenden Rezeption der Rückmeldungen nicht genügend Aufmerksamkeit gewidmet wurde: „Schließlich wird mit den an Schulrückmeldungen verbundenen Erwartungen unterstellt, dass die dargestellten Ergebnisse in professioneller und problemloser Weise rezipiert und verarbeitet werden“ (Watermannn, et al., 2003, S. 92).

Wie sollten Ergebnisrückmeldungen beschaffen sein, um für die Lehrkräfte verständlich und nützlich sein zu können? Shirley äußert hierzu: „Not surprisingly, teachers have indicated

that more iterative formative forms of assessment that allow them to respond with instructional modifications immediately are far more important than summative assessments with results that are returned to teachers far too late to have any pedagogical utility“ (Shirley, 2008, S. 37). Schneewind (2007) berichtet ähnliche Befunde und konstatiert, dass sich Ergebnisrückmeldungen aus Sicht der Lehrkräfte unter anderem durch die folgenden Merkmale auszeichnen sollten. Sie sollten

- zeitnah erfolgen,
- leicht zu erhalten sein,
- einfach zu handhaben und zu verstehen sein,
- keine statistischen Grundkenntnisse voraussetzen,
- nicht zu viele Informationen beinhalten und
- Hinweise auf Verwendungsmöglichkeiten geben, also die Möglichkeiten und Grenzen der Informationen thematisieren (vgl. Schneewind, 2007).

Der letzte Punkt verweist bereits auf den Aspekt der Nutzbarmachung der Rückmeldung für die unterrichtliche Praxis. Grundlegend für die Auseinandersetzung mit Ergebnisrückmeldungen und ihre Übersetzung in zielführendes pädagogisches Handeln ist die Verfügbarkeit von Bezugsnormen. Neben sozialen Bezugsnormen auf Schul- oder Bundeslandebene in so genannten „fairen Vergleichen“ kann dies auch eine kriteriale Bezugsnorm, wie etwa die Niveaus eines Kompetenzstufenmodells umfassen (vgl. bspw. U. Koch et al., 2006; Watermann et al., 2003).

Das zentrale Problem der Nutzung von Ergebnisrückmeldungen für die Unterrichtsgestaltung ist allerdings die Tatsache, dass Lehrkräfte oftmals ihre Schülerinnen und Schüler und nicht sich selbst als Adressaten der Rückmeldung verstehen. Die Schul- und Unterrichtsgestaltung scheitern dann nicht an Akzeptanzproblemen oder mangelhafter Verständlichkeit, sondern an der Transferleistung von den Leistungsergebnissen der Schülerschaft zum Unterrichtshandeln der Lehrkräfte. Das bedeutet, dass die in den Rückmeldungen enthaltenen Informationen zwar oftmals oberflächlich verstanden werden, aber nicht in unterrichtsrelevante Handlungen übersetzt werden können. So werten die Lehrkräfte Ergebnisrückmeldungen zumeist ausschließlich als Information hinsichtlich der Leistungsstände der Schülerinnen und Schüler, stellen aber keinen Zusammenhang zwischen diesen Leistungsständen und ihrem eigenen pädagogischen Handeln oder ihrer Unterrichtsgestaltung her. Dementsprechend führen Ergebnisrückmeldungen zu Schulleistungstests selten dazu, dass die Lehrkräfte die Rückmeldung als Aktionsimpuls verstehen und beispielsweise ihre

Unterrichtspraxis verändern (vgl. Schrader & Helmke, 2004; Schneewind, 2007). Schneewind (2007) kommt somit zu dem folgenden Fazit: „Die Akzeptanz der Ergebnisrückmeldungen ist oftmals hoch, die Bereitschaft zur reflektorischen Auseinandersetzung besteht, jedoch die Umsetzung der so gewonnenen Erkenntnisse in Handlungen, Unterrichts- oder Schulentwicklung ist wenig ausgeprägt“ (S. 236).

Dass aber auch eine Rückmeldung von erzielten Leistungen an die Schülerinnen und Schüler starke positive Wirkungen entfalten kann, belegen Hattie, Biggs und Purdie (1996) in einer großen metaanalytischen Studie, in der sie die Effekte verschiedenster Maßnahmen und Kontextfaktoren auf das schulische Lernen untersuchen. Hierbei finden sie einen Effekt in Höhe von *d=0.65* für Feedback beziehungsweise Rückmeldung der in Tests erzielten Leistungen auf das schulische Lernen (S. 115).

Insgesamt kann festgehalten werden, dass die Nutzbarmachung von Ergebnisrückmeldungen auf Seiten der Lehrkräfte selten an Akzeptanzproblemen oder mangelhafter Verständlichkeit scheitert. Obwohl in diesem Kontext auch im Auge behalten werden muss, dass es keinesfalls zu einer testbedingten Überlastung der Schulen sowie der Schülerinnen und Schüler – und damit zu einer Verringerung der Akzeptanz – kommen darf, ordnet sich die Frage nach Anzahl und Umfang von Tests dem zentralen Problem unter, in welcher Weise die in den Studien gewonnen Ergebnisse nutzbar gemacht werden können. Auch in dieser Hinsicht lohnt ein Blick in Staaten, die bereits über einen reicheren Erfahrungsschatz verfügen (vgl. van Ackeren, 2007; Oelkers & Reusser, 2008). In verschiedenen europäischen Staaten ist versucht worden, die Ergebnisse von Schulleistungsstudien marktorientiert zu verwenden und die Verteilung finanzieller Mittel an das Erreichen beziehungsweise Verfehlen bestimmter Vorgaben zu knüpfen. Die Übertragung marktähnlicher Strukturen auf das Bildungssystem führt so unter anderem dazu, dass die in Einzelschulen erzielten Ergebnisse öffentlich zugänglich gemacht und den Eltern verstärkt Schulwahlmöglichkeiten eingeräumt werden (vgl. van Ackeren, 2007). Internationale Erfahrungen lassen ein solches öffentliches Ranking jedoch nicht sinnvoll erscheinen. In Großbritannien, wo regelmäßig leistungsbezogene Rangreihen publiziert werden, in denen explizit erfolgreiche und eher schlecht abschneidende Schulen benannt werden, hat man erkannt, dass ein derartiges Vorgehen maßgeblichen Einfluss auf die Schulen und deren öffentliche Wahrnehmung hat (vgl. van Ackeren, 2007).

Zielführender als ein solches Vorgehen scheint eine systematische Analyse der empirischen Daten, auf deren Grundlage maßgeschneiderte Schritte im Sinne einer nachhaltigen Schulentwicklung vor Ort eingeleitet sowie entsprechende Unterstützungs- und Beratungsangebote etabliert werden können. Entscheidend ist und bleibt, dass die Rückmeldungen so gestaltet sind, dass sie für die Lehrkräfte konkrete Handlungsoptionen

eröffnen und der Transfer der berichteten Ergebnisse in zielgenaue Förderung der Schülerinnen und Schüler und die optimierte Gestaltung des Unterrichts überführt werden können. So formulieren auch Vertreter des EMSE-Netzwerks:

> Nach vorliegenden Erfahrungen wird der Zugang der Lehrerschaft zu Verfahren datengestützter Schul- und Unterrichtsentwicklung durch möglichst konkrete Informationen zum Leistungsstand der Lernenden, durch Verknüpfung mit bereits vorliegenden diagnostischen Informationen seitens der Schulen und durch Hinweise auf Möglichkeiten der innerfachlichen Weiterarbeit begünstigt. (EMSE-Netzwerk, 2008, S. 4)

Auch Tippelt und Schmidt benennen die handlungsrelevante Aufbereitung und Rückmeldung empirischer Befunde an die Lehrkräfte als eine der zentralen, aber bislang nicht hinreichend gelösten Herausforderungen der empirischen Bildungsforschung und fragen: „Wie können die analytischen Ansätze und Ergebnisse der Bildungsforschung handlungsorientiert an die Träger von Bildungsentscheidungen und das verantwortliche pädagogische Personal vermittelt werden?“ (Tippelt & Schmidt, 2010, S. 15). Hierauf eine befriedigende Antwort zu finden ist aus meiner Sicht eine der zentralen Herausforderungen der interdisziplinären Zusammenarbeit zwischen Fachdidaktik und pädagogisch-psychologischer Diagnostik sowie eines der entscheidenden Forschungsdesiderate der aktuellen empirischen Bildungsforschung.

4.4 Teaching to the Test[2]

Die regelmäßige Durchführung standardisierter Testverfahren kann Rückwirkungen auf das schulische Lehren und Lernen haben. Dieser Effekt wird in der einschlägigen Literatur als Washback bezeichnet (vgl. bspw. Cheng, Watanabe & Curtis, 2004).

Ein zentrales Diskussionsthema, welches in diesem Kontext wiederkehrend thematisiert wird, ist das Problem des *Teaching to the Test*. Diese Begrifflichkeit zielt auf die Tatsache ab, dass mitunter vor Testungen insbesondere diejenigen Inhalte unterrichtet und gezielt solche Aufgabentypen eingeübt werden, die in den Tests eine entscheidende Rolle spielen werden. Überdies kann eine Verschiebung der erteilten Unterrichtsstunden zugunsten jener Fächer erfolgen, die in den anstehenden Tests abgeprüft werden (vgl. Stecher & Barron, 1999). Solche Testvorbereitungen können das Ausmaß von systematischem, massivem Training annehmen, hinter dem die Zielstellung eines kompetenzorientierten, interessanten und motivierenden

[2] Der Abschnitt „4.4 Teaching to the Test“ basiert teilweise auf Textpassagen des Beitrags: Böhme, K. (2006). Testen: ja – Den Unterricht verarmen: nein. *Grundschule, 5*, 8-10.

Unterrichts zurücktritt. Derartige Erfahrungsberichte sind insbesondere aus den USA nach Verabschiedung des *No Child Left Behind Act* bekannt geworden, wo die Verarmung des Unterrichts, eine Vernachlässigung der nicht in den Tests abgeprüften Fächer und eine abnehmende Motivation der Schülerinnen und Schüler durch intensive Vorbereitungen auf anstehende Leistungserhebungen wiederholt beklagt worden sind. Shirley äußert in diesem Zusammenhang:

> Surveys indicate that only 15% of American teachers believe that the federal No Child Left Behind Act of 2001 mandating the testing of all children in grades 3 – 8 and prior to graduation from high school is improving American education (Public Agenda 2006). 44% of school districts report that they have narrowed their curricula to increase time on tested subjects, reducing their offerings especially in the areas of music, art, and foreign languages (Center on Education Policy 2007). (Shirley, 2008, S. 36)

Als ursächlich für dieses intensive Teaching to the Test kann der Tatbestand diskutiert werden, dass in den USA die in den Tests erzielten Ergebnisse nicht nur für die Schulen, sondern zum Teil auch für die einzelnen Lehrkräfte weitreichende Konsequenzen haben, sodass aufseiten der Schulleitungen enormes Interesse an einem guten Abschneiden in den entsprechenden Leistungserhebungen besteht (vgl. Köller, 2010). Eine solche Praxis, bei der die Vergabe von Belohnungen und Sanktionen an Testergebnisse gekoppelt ist, wird als *High Stakes Testing* bezeichnet. In aller Regel sind die Konsequenzen der individuellen Testbearbeitung hierbei primär für das jeweilige Individuum relevant, so wird die Studienzulassung in den USA beispielsweise über die erfolgreiche Bearbeitung des SAT (vgl. CollegeBoard, 2011) geregelt. Im hier relevanten Kontext werden allerdings die Schulen für schlechte Leistungen ihrer Schülerinnen und Schüler in flächendeckenden Vergleichsarbeiten sanktioniert, was als zusätzliches Problem gewertet werden kann.

Dass eine solche Nutzung standardisierter Testverfahren zu empirischen Befunden führt, die für die Schulentwicklung unbrauchbar, da unzuverlässig sind, belegt das Phänomen der *Inflating Gain Scores*. Dieses bezeichnet den Umstand, dass beträchtliche Diskrepanzen zwischen den Ergebnissen des National Assessment of Educational Progress (NAEP), also dem stichprobenbasierten Bildungsmonitoring in den Vereinigten Staaten, und verschiedenen, bundesstaatenspezifischen Vergleichsarbeiten gefunden wurden. Während die Schülerleistungen in den flächendeckenden Vergleichsarbeiten (High Stakes Tests) kontinuierlich ansteigen, finden sich in NAEP (Low Stakes Test) keine entsprechenden Leistungszuwächse (vgl. bspw. Koretz, 2002, 2005; Green, Winters & Forster, 2003).

Andere Staaten berichten jedoch in wesentlich geringerem Ausmaß über negative Erfahrungen mit Teaching to the Test. So scheinen beispielsweise in Schweden die nationalen Leistungserhebungen nicht dazu zu führen, dass der Unterricht auf die Vermittlung von einfach abfragbarem Wissen reduziert wird. Allerdings wurden auch von schwedischen Lehrerinnen und Lehrern dahingehende Bedenken geäußert, dass sich eine Schwerpunktsetzung zugunsten derjenigen Unterrichtsfächer andeutet, die in den nationalen Tests abgeprüft werden (Stanat, 2002). In den Niederlanden wird das Problem des Teaching to the Test dadurch reduziert, dass in den Testungen vorrangig solche Aufgaben eingesetzt werden, die einen Einsatz von Kenntnissen und Fertigkeiten in anwendungsbezogenen Kontexten erfordern (Stanat, 2002). Wird nun der Unterricht an die kognitiven Anforderungen solcher Testaufgaben angepasst, dann stellt dies sogar einen wünschenswerten Schritt im Sinne einer zunehmenden Kompetenzorientierung des Unterrichts dar. Köller (2010) bezeichnet eine solche Form der Testvorbereitung als *Coaching* und führt für den deutschsprachigen Raum an, dass es erstrebenswert wäre, wenn die auf Sprachhandlungskompetenz abzielenden Bildungsstandards für die erste Fremdsprache und die hierauf bezogenen Tests dazu führen würden, „mehr Unterrichtszeit für kommunikative Fertigkeiten in den Bereichen des Hörens und Sprechens“ (S. 545) zu verwenden.

Neben den etwaigen negativen Auswirkungen eines massiven Vorbereitens auf Leistungserhebungen in Bezug auf die Unterrichtsqualität liegt das Problem auch auf der Ebene der Interpretierbarkeit der Ergebnisse der fraglichen Schulleistungsstudien. Sind die erzielten Ergebnisse vorrangig das Resultat kurzfristigen, massiven und gezielten Trainings, so kann nicht mehr die eigentliche Kompetenz der Schülerinnen und Schüler diagnostiziert werden, sondern lediglich das Ausmaß an Vorbereitung. So entstandene Ergebnisse könnten nicht als aussagekräftige Vergleichsgrundlage herangezogen werden und würden auch keine zielführende Interpretation der empirischen Daten bzgl. der in Frage stehenden Qualitätsentwicklung des deutschen Bildungssystems gestatten. Eine Vorbereitung der Schülerinnen und Schüler in Form eines unreflektierten, wiederholenden Einübens von Testaufgaben kann somit weder für stichprobenbasierte, internationale oder nationale Schulleistungsstudien noch für jährliche, flächendeckende Vergleichsarbeiten als sinnvoll und unterstützenswert angesehen werden. Bestrebungen von Seiten politischer Verantwortungsträger, deren primäres Interesse ein gutes Abschneiden des durch sie vertretenen Bundeslandes ist und die aus diesem Grund mitunter ein gezieltes Einüben befürworten und empfehlen, müssen in diesem Sinne kritisch hinterfragt werden.

Allerdings darf an dieser Stelle nicht unerwähnt bleiben, dass eine gewisse *Vorbereitung auf die Testsituation* tatsächlich nützlich und sinnvoll sein kann. Da insbesondere Schülerinnen und Schüler des Primarbereichs im Verlauf ihres Schulbesuchs zumeist nur wenig Erfahrungen mit

den in standardisierten Leistungstests eingesetzten Formaten und der standardisierten Weise der Instruktion der Aufgabenbearbeitung gesammelt haben, kann die Testsituation aufgrund ihrer Neu- und Andersartigkeit bei den Kindern zu Verunsicherung führen. Geringe Testleistungen könnten dann zu einem gewissen Grad auf die fehlende Vertrautheit mit Aufgabenformaten und standardisierter Instruktion und nicht notwendigerweise auf mangelnde Kompetenz zurückgeführt werden. Um dies auszuschließen, kann es also zweckdienlich sein, den Schülerinnen und Schülern bereits im Vorfeld zu erklären, dass eine Testsituation anders beschaffen ist als das alltägliche Unterrichtsgeschehen und die Lehrkraft während des Tests nicht in der gewohnten Weise Hilfe und Unterstützung anbieten kann. Auch sind die in standardisierten Leistungstests häufig eingesetzten geschlossenen Aufgabenformate – und hier insbesondere Multiple-Choice-Items – für viele Schülerinnen und Schüler nach wie vor ungewohnt, so dass sie hier möglicherweise nicht aufgrund mangelnder Kompetenz, sondern aufgrund eines Unverständnisses des Aufgabentyps an der korrekten Bearbeitung scheitern. Testvorbereitung muss also nicht die mit negativen Konsequenzen beladene Form des Teaching to the Test annehmen, sondern kann auch die Gestalt eines Heranführens an eine ungewohnte Situation annehmen. Dass eine solche Vertrautheit mit der Testsituation und den zu bearbeitenden Aufgabentypen zu einer Verbesserung der Testleistung führen kann, liegt nahe. Da diese Verbesserung aber darauf basiert, dass kognitive Kapazitäten nicht an falscher Stelle durch *formale Aspekte* gebunden werden, sondern die gesamte Aufmerksamkeit auf die *inhaltlichen Aspekte* der Instruktionen und Aufgaben gelenkt werden kann, sollte diese Form der Vorbereitung nicht als eine die wahren Leistungen verfälschende Manipulation fehlinterpretiert werden. In diesem Sinne eröffnet also eine Vorbereitung auf die im Test zu erwartenden Instruktionen und Aufgabentypen den Schülerinnen und Schülern die Chance, ihre tatsächlich vorhandenen Kompetenzen auch einzusetzen und ist somit zweckdienlich, da sie Gleichheit in den Voraussetzungen herstellt und Unterschiede im Bearbeitungserfolg mit größerer Sicherheit auf Unterschiede in den Kompetenzständen zurückführbar sind. Um ein solches Vertrautmachen mit den möglichen Instruktionen und Aufgabenstellungen in bevorstehenden Tests zu ermöglichen, scheint die Veröffentlichung von exemplarischem Testmaterial ein geeigneter Weg zu sein, der nicht zuletzt auch die Transparenz und somit die Akzeptanz der Schulleistungsstudien fördert.

Studien, die sich mit der Frage beschäftigen, welche Faktoren eine Auswirkung auf die Lernfähigkeit und Leistungen von Schülerinnen und Schülern haben, untersuchen auch, welche Zusammenhänge zwischen diesen Variablen und Testungen sowie der Vorbereitung auf diese Tests bestehen. Hattie, Biggs und Purdie (1996) berichten in einer Meta-Analyse, in die 51 Studien zu diesem Thema einbezogen wurden, dass durch die Durchführung von Testungen im Mittel ein Effekt von .30 auf die Lernfähigkeit beobachtet wird (Hattie, Biggs & Purdie, 1996,

S. 115). Hinsichtlich des Phänomens des Teaching to the Test merken die Autoren an: „Feuerstein's (1969) Instrumental Enrichment program, produced strong effect sizes on performance (0.69) […] A detailed examination of the training activities, […] raises the suspicion that again what we are seeing here is something very akin to 'teaching to the test' " (Hattie, Biggs & Purdie, 1996, S. 121). Es wurden also große positive Effekte auf die Schülerleistungen beobachtet, die in einem Lernprogramm erzielt wurden, das hohe Ähnlichkeit mit Teaching to the Test aufwies. Diese positiven Effekte beschränkten sich allerdings auf die Schülerleistungen und konnten nicht in gleicher Weise für die Lernfähigkeit und die emotionale Verfassung der Schülerinnen und Schüler nachgewiesen werden.

In Bezug auf das Phänomen des Teaching to the Test sollte abschließend festgehalten werden, dass der Grad, in dem dieses zum Problem werden kann, in entscheidender Weise durch die Konzeption der verwendeten Testaufgaben bestimmt wird. Solange es gelingt, verständnisorientierte Tests zu entwickeln, in denen nicht primär Wissen abgefragt wird, sondern die in erster Linie darauf abzielen, erworbene Kompetenzen in neuartigen Kontexten flexibel anzuwenden, wird der Unterricht so gestaltet werden müssen, dass die Inhalte von den Schülerinnen und Schülern tatsächlich verstanden werden. Ein massives, systematisches Trainieren und Üben bestimmter Aufgabentypen kann die Leistung in derartigen Tests kaum verbessern. Stanat (2002) weist darauf hin, dass ein so geartetes Teaching to the Test sogar wünschenswert erscheint, da es zum Ziel hätte, den Erwerb grundlegender Kompetenzen zu fördern (vgl. hierzu auch Blum, 2006).

4.5 Zukunftsperspektiven der Erziehungswissenschaft

Wie in Abschnitt 2.1.3 angedeutet, gestaltet sich das Verhältnis der Erziehungswissenschaft zur empirischen Bildungsforschung problematisch. Dies betrifft zum einen das grundlegende wissenschaftliche Selbstverständnis und die als adäquat erachteten Forschungsmethoden sowie zentrale Inhalte des Faches und zum anderen die Frage der Zukunftsperspektive. Beide Aspekte sind eng miteinander verflochten.

Pädagogische Empiriekritik in ihrer fundamentalsten Form bezieht sich auf die Aussage, dass empirische Verfahren das „Pädagogisch-Eigentliche" (Heid, 1996, S. 2) verfehlen würden, da pädagogisch bedeutsame Gegebenheiten wie beispielsweise Bildsamkeit oder Mündigkeit sich der unmittelbaren Beobachtung entzögen und keine „bloßen Tatsachen" (Schurr, 1975, S. 6) oder „empirischen Faktizitäten" (Schmied-Kowarzik, 1993, S. 90) darstellen würden. Insgesamt kann man sich allerdings nicht des Eindrucks erwehren, dass Empiriekritik sich mitunter eher aus der *Unerwünschtheit* erziehungswissenschaftlicher Empirie als aus ihrer *Unmöglichkeit* speist. Heid

(1996) formuliert in diesem Zusammenhang pointiert, dass offen bleibe „woher Empiriekritiker eigentlich etwas über jene nicht (unmittelbar) beobachtbaren Gegenstände wissen, über die sie sehr ausführlich reden oder schreiben, und nach welchem Verfahren sie überprüfen, ob das, was sie über diese Gegenstände aussagen, zutrifft oder nicht“ (S. 7).

Die Abwehrhaltung von Vertretern der Allgemeinen Erziehungswissenschaft, die sich dem Verständnis der traditionellen Pädagogik verpflichtet fühlen, gegenüber der empirischen Bildungsforschung erschöpft sich allerdings nicht in empiriekritischen Aussagen, sondern gipfelt gelegentlich im Postulat einer „feindlichen Übernahme“ der Erziehungswissenschaft durch die empirische Bildungsforschung beziehungsweise durch die Pädagogische Psychologie (vgl. bspw. Gruschka 2006a, 2006b). Es wird beklagt, dass die empirische Bildungsforschung innerhalb der – beziehungsweise in Konkurrenz zur – Erziehungswissenschaft immer stärker in den Vordergrund trete und diese Bevorzugung zu Lasten anderer erziehungswissenschaftlicher Fragestellungen und Forschungsrichtungen gehe. Aus Sicht der Erziehungswissenschaft bestehen in der heutigen empirischen Bildungsforschung thematische Leerstellen, welche auf eine einseitige Orientierung an den Standards pädagogisch-psychologischer Forschung zurückzuführen sind (vgl. Merkens, 2006, S. 14). Nach Jendrowiak (1998) ist die Erziehungswissenschaft eine Disziplin, die sich originär über ihre Gegenstände (Bildung, Erziehung, Lernen, Sozialisation usw.) bestimmt. Zur aktuellen Situation der Erziehungswissenschaft und ihren Gegenständen bemerkt er: „Wenn die Pädagogen aber die Gegenstände ihrer eigenen Disziplin in Frage stellen und sich die Nachbardisziplinen, wie z. B. die Psychologie […] immer mehr der originären pädagogischen Gegenstände bemächtigen; wenn die Pädagogik ihre Gegenstände abtritt, verliert sie ihre Basis und wird sinnlos“ (Jendrowiak, 2001, S. 2). Neben dieser wahrgenommenen Verdrängung durch „rivalisierende“ Disziplinen und inhaltlichen Verkürzung pädagogischer Inhalte kommen Probleme der Verteilungsgerechtigkeit zum Tragen, welche durch das massive Wachstum der empirischen Bildungsforschung in den letzten Jahren befördert werden. Die Verengung der Disziplin zeige sich insbesondere in der Umwidmung beziehungsweise Neubesetzung von Professuren und anderen Personalstellen (vgl. bspw. Ruhloff, 2007; Zedler & Döbert, 2010). Dass diese wiederholt vorgetragene Klage allerdings kein wissenschaftliches Argument, sondern ein „interessengeleitetes, politisches Statement“ der Kritiker der empirischen Bildungsforschung darstellt, betont beispielsweise Klieme (2007, S. 141).

Was in diesem Zusammenhang in besonderem Maße erstaunt, ist allerdings der Umstand, dass einige Vertreter der Erziehungswissenschaft es nachdrücklich ablehnen, durch ihre wissenschaftliche Forschung für die Gesellschaft nützlich zu sein. In gewohnt bissiger Weise kommentiert Gruschka (2006a) die Fremdsicht auf die Erziehungswissenschaft:

> […] was man andernorts, dort also, wo die Steuerung des Wissenschaftssystems inzwischen faktisch erfolgt, von der akademischen Pädagogik erwartet und von ihrer bisherigen Leuchtturmdisziplin, der Bildungstheorie hält: Weil diese sich nicht um die Aufklärung und die Optimierung der realen Verhältnisse in der Schule kümmere, […], sei diese überflüssig, abzuschaffen und durch eine Wissenschaft zu ersetzen, die als ‚international visible normal science', der Praxis und Politik das Wissen zur Verfügung stellt, mit dem pädagogische Handlungsmuster in der öffentlichen Erziehung verändert und verbessert werden können. (Gruschka, 2006a, S. 7)

Auch Hansel (2007) spricht davon, dass sich die Erziehungswissenschaft „auf die verhängnisvolle Suche nach dem eigenen Nutzen" (S. 194) eingelassen habe und Gefahr laufe, ihre Unabhängigkeit zu verlieren. Man müsse nun alles daran setzen, dieser „Brauchbarkeitstendenz" mit der „eigenen definitorischen Kraft" entgegenzuwirken, um sich nicht selbst zur „Reparaturwissenschaft" zu degradieren (Hansel, 2007, S. 194). Solche Äußerungen überraschen in hohem Maße und lassen die Frage aufkommen, ob Erziehungswissenschaftler tatsächlich für sich in Anspruch nehmen, ihre Wissenschaft sei Selbstzweck. Auch widerspricht eine solche Verweigerungshaltung dem normativen Charakter der Disziplin, die selbstverständlich für sich in Anspruch nimmt, zu definieren, wie gelungene Erziehung und Bildung aussehen. Dies zu tun, ohne sich daran beteiligen zu wollen, zu überprüfen, ob es sich hierbei auch um gelingende Vorstellungen handelt und – falls nein – Wege aufzuzeigen, wie die Zielvorstellung gesamtgesellschaftlich erreicht werden kann, erscheint mir wenig zukunftsträchtig.

Nicht zuletzt aufgrund der häufig polemisch geführten Diskussionen um die Sinnhaftigkeit empirischer Forschung im Bildungsbereich und der hiermit einhergehenden Zurückweisung steht die traditionsorientierte Erziehungswissenschaft in den letzten Jahren unter einem gewissen Legitimationsdruck (vgl. Hansel, 2007). Diese Rechtfertigungsnot bezieht sich zum einen auf öffentlich laut gewordene Stimmen, die Erziehungswissenschaft sei „nur bedingt wissenschaftlich" (vgl. bspw. Kahl & Spiewak, 2005; Weiler, 2002) und zum anderen auf das Problem, dass die Erziehungswissenschaft hinsichtlich ihrer eigenen Gegenstände keine befriedigenden Antworten auf gesellschaftlich dringende Fragen liefern konnte – oder wollte (s. o. Gruschka, 2006a). Beide Punkte sollen an dieser Stelle kurz ausgeführt werden.

Der Vorwurf, dass die Erziehungswissenschaft den Kriterien wissenschaftlichen Arbeitens nicht vollständig genügen kann, wurde in dieser Deutlichkeit zu Beginn dieses Jahrtausends zuerst von Weiler (2002, 2003) formuliert. In der Folge wurden der Erziehungswissenschaft verschiedene Mängel hinsichtlich den international anerkannten Standards wissenschaftlichen Arbeitens vorgeworfen, was nicht nur innerhalb der

wissenschaftlichen Gemeinschaft thematisiert, sondern auch in den meinungsbildenden Printmedien öffentlich diskutiert wurde (vgl. bspw. Kahl & Spiewak, 2005; Weiler, 2002).

Der Stand der wissenschaftlichen Forschungstätigkeit in der Erziehungswissenschaft wurde von Hornbostel und Keiner (2002) begutachtet. Hier zeigte sich, dass beispielsweise die Anzahl der Publikationen sowie das Drittmittelvolumen je Wissenschaftler in der Erziehungswissenschaft geringer als in anderen, vergleichbaren Disziplinen ausfallen. Auch ist nur ein geringer Teil der forschungsfördernden Drittmittel DFG finanziert. Der Anteil von Landesmitteln, für die in der Regel keine gutachterlichen Stellungnahmen wie bei DFG-Anträgen eingeholt werden, ist vergleichsweise hoch. Publikationen in Fachzeitschriften mit Peer-Review sind eher selten, einen hohen Anteil an den selbstberichteten Veröffentlichungen machen Unterrichtsmaterialien und anwendungsorientierte, didaktische Grundlageninformationen aus, bei denen es sich also nicht um die Publikation von Forschungsergebnissen beziehungsweise um Fachbeiträge im eigentlichen Sinne handelt (vgl. zu allen genannten Befunden Hornbostel & Keiner, 2002). Diese Befundlage begründe sich allerdings auch durch die reflexionsorientierte, geisteswissenschaftliche Tradition der Disziplin, durch die eine Anwendung der genannten Kriterien zu Ungunsten der Erziehungswissenschaft ausfalle (Hornbostel & Keiner, 2002; Merkens, 2005). Aus wissenschaftlicher Perspektive schätzt Prenzel (2006) die methodischen Standards der empirischen Bildungsstandards als hoch ein und lobt die sich hier abzeichnende positive Tendenz, gleichzeitig räumt er ein, dass „innerhalb der Erziehungswissenschaft die Entwicklung grundlegender methodischer Standards etwas langsam vorankommt […]“ (Prenzel, 2006, S. 76).

Zusammenfassend formuliert Weiler das zwiespältige Verhältnis der Erziehungswissenschaft zu quantitativ-empirischem Arbeiten und ihre wissenschaftstheoretische Grundposition wie folgt:

> Im Falle der deutschen Erziehungswissenschaft zeigt sich diese Ambivalenz in der Neigung, der geistesgeschichtlichen, textkritischen und wertphilosophischen Reflexion über Erziehung den Vorzug vor der empirischen Forschung, und der normativen Aussage den Vorzug vor der analytischen Beschreibung und Erklärung von Sachverhalten zu geben. (Weiler, 2003, S. 194f.)

Eine selbstkritische Auseinandersetzung mit den hinsichtlich der wissenschaftlichen Arbeitsweise identifizierten Problemen findet in der Disziplin vereinzelt statt (vgl. bspw. Hornbostel & Keiner, 2002; König & Zedler, 2004). Auch Brezinka (2004, S. 29f.) weist auf die „Wucherung metapädagogischer Reflexionen auf Kosten pädagogischer Erkenntnisse“ hin und äußert, dass eine übertriebene Zuwendung zu komplizierten wissenschaftsphilosophischen Fragen die

pädagogische Forschung gehemmt hat. Ferner unternimmt die Erziehungswissenschaft in den letzten Jahren vermehrt Anstrengungen, den Leistungserwartungen der Öffentlichkeit gerecht zu werden, ihre Stärken und Schwächen in Datenreports zu dokumentieren (vgl. zuletzt Tillmann, Rauschenbach, Tippelt & Weishaupt, 2008) und sich dem wissenschaftspolitischen Diskurs zu öffnen (vgl. Merkens, 2005).

Gelegentlich werden die genannten Kritikpunkte aber auch als anmaßend und nicht berechtigt zurückgewiesen (vgl. bspw. Gruschka, 2006a, 2006b; Merkens, 2006). Vereinzelt findet eine Reaktion auf kritische Stellungnahmen zur aktuellen wissenschaftlichen Leistungsfähigkeit der (Allgemeinen) Erziehungswissenschaft sowie dem wachsenden Einfluss empirischer Bildungsforschung nur in polemisch überspitzter Form statt, die einen wissenschaftlichen Dialog erschwert, wenn nicht ausschließt (vgl. bspw. Gruschka, 2006a, 2006b, 2007; Jahnke, 2008; Ruhloff, 2007).

Dass die Erziehungswissenschaft die an sie gestellten Erwartungen teilweise nicht erfüllen kann, wurde insbesondere hinsichtlich der Lehrerbildung thematisiert. Durch die von Seiten der empirischen Bildungsforschung vorgelegten Ergebnisse wurde auch für die Vertreter der Erziehungswissenschaft deutlich, dass in der Lehreraus- und -weiterbildung Defizite vorliegen, die durch die Erziehungswissenschaft selbst weder erkannt wurden, noch beseitigt werden konnten. Nach Ansicht von König und Zedler (2004, S. 77f.) kann die Allgemeine Pädagogik den gesellschaftlichen Erwartungen hier nicht gerecht werden. Konkret heiße dies, dass für die praktisch tätigen Pädagogen das erhoffte konkrete Handlungswissen – trotz des gesellschaftlichen Bedarfs – durch die Allgemeine Pädagogik nicht zur Verfügung gestellt werde. Auch Hansel (2007) räumt – sehr vorsichtig, aber unmissverständlich – ein, dass „[…] der Einwand, dass erziehungswissenschaftliche Forschung für die erzieherischen Handlungsprobleme in der Schule und im Unterricht nur sehr bedingt Hilfe anbietet, zumindest teilweise nicht leicht zu entkräften“ (S. 190) sei. Auch der Wissenschaftsrat beschreibt die Rolle der Erziehungswissenschaft hinsichtlich der Lehrerbildung als vernachlässigbar: „Ihr ehemals ausgeprägter Bezug auf die Lehrerbildung hat an Verbindlichkeit verloren oder wurde […] vollends aufgegeben“ (Wissenschaftsrat, 2001, S. 23 f.). Die von Seiten der Erziehungswissenschaft hinsichtlich dieser Missstände vorgetragene Rechtfertigung bezieht sich zumeist auf die Tatsache, dass die Lehrerausbildung – auch wenn es sich hierbei um einen pädagogischen Beruf handelt – in überwiegendem Umfang ein Fach- und kein erziehungswissenschaftliches Studium sei (vgl. Merkens, 2006; Terhart, 2003).

Während Empiriekritiker nach wie vor grundsätzliche Zweifel äußern, ob Pädagogik überhaupt als empirische Wissenschaft betrieben werde könne (vgl. Heid, 1996), gehen andere Vertreter davon aus, dass die Erziehungswissenschaft mittlerweile bereits eine empirisch arbeitende Sozialwissenschaft und die klassische, philosophisch ausgerichtete Pädagogik „am Ende ihrer Epoche" angelangt ist (König & Zedler, 2004, S. 81; vgl. auch Sailer, 2007).

Vor diesem Hintergrund stellt sich die Frage nach der Zukunft der Erziehungswissenschaft. Diese kann in verschiedene Richtungen weisen.

Zum einen ist denkbar, dass traditionelle Inhalte bewahrt werden und es der Allgemeinen Erziehungswissenschaft gelingt, sich dauerhaft neben der empirischen Bildungsforschung zu etablieren. In einer solchen Zukunftsperspektive wird die traditionelle Erziehungswissenschaft durch den Aufstieg der empirischen Bildungsforschung nicht obsolet, sondern stellt eine Ergänzung dar. Hansel (2007) thematisiert in diesem Sinne wissenschaftliche Anliegen, welche sich für die Erziehungswissenschaft aus der empirischen Bildungsforschung ergeben können:

> Zwar ist sie [die empirische Bildungsforschung] selbst als Partikulardisziplin der Erziehungswissenschaft nicht in der Lage, durchtragende Erziehungs- und Bildungskonzepte zu entwickeln und theoretisch zu begründen, aber sie stellt für den Fortgang der erziehungswissenschaftlichen Grundsatzdebatte vergleichende Daten zur Verfügung, sie ersetzt diese Debatte jedoch nicht. Das ist ein positives Signal, das aufgenommen werden muss – freilich unaufgeregt, undogmatisch, unabhängig […]. Hansel (2007, S. 195 f.)

Weiter heißt es bei Hansel zum Stellenwert der Erziehungswissenschaft:

> Die Erziehungswissenschaft ist nicht die einzige Wissenschaft, die einen Beitrag zu Entscheidungen leistet, die im Zusammenhang mit Erziehung, Bildung, Unterricht usw. stehen. […] und die Erziehungswissenschaft ist gut beraten, nicht dort Zuständigkeit zu reklamieren, wo andere Wissenschaften weitaus entwickelter sind als sie selbst. (Hansel, 2007, S. 195)

Dies kann so verstanden werden, dass die Erziehungswissenschaft als Disziplin entsprechend ihren tradierten Gegenständen nach Räumen suchen sollte, die spezifisch erziehungswissenschaftlich bearbeitet werden können und sollten. Diese Perspektive birgt aus meiner Sicht allerdings die Gefahr der Abgrenzung und Isolation, da die Rolle der empirischen Forschung in Bildungsfragen in ihrer Bedeutung in den nächsten Jahren und Jahrzehnten vermutlich nicht nachlassen wird.

Zum anderen ist eine Zukunft der Erziehungswissenschaft denkbar, in der sich diese der empirischen Bildungsforschung deutlich stärker zuwendet, sich aktiver auf diesem Forschungsgebiet betätigt und hierbei ihre Stärken einbringt. Tatsächlich hat sich der Anteil *erziehungswissenschaftlicher* Bildungsforschung in den letzten Jahren bereits deutlich erhöht (vgl. Zedler & Döbert, 2010). Voraussetzung für eine solche Entwicklung ist allerdings, dass sich die Erziehungswissenschaft dem hierfür erforderlichen Methodenrepertoire öffnet und sich dieses zu eigen macht. Prenzel (2006) äußert in diesem Zusammenhang:

> Tatsächlich dürften die erziehungswissenschaftlichen Beiträge zur Bildungsforschung besser sichtbar werden, wenn sich mehr Kolleginnen und Kollegen aus der Erziehungswissenschaft mit einer ausgeprägten Forschungsorientierung und mit einem anspruchsvollen Methodenbewusstsein in diesem Feld engagieren würden. Dass dabei eine Orientierung am internationalen Erkenntnisstand unabdingbar ist, liegt ebenso auf der Hand wie die Notwendigkeit eines empirisch-analytischen Herangehens. (Prenzel, 2006, S. 76)

In diesem Zusammenhang ist der Umstand bedeutsam, dass sich das Erfordernis und der Wunsch nach Anwendung quantitativ empirischer Methoden in der Erziehungswissenschaft momentan schneller entwickelt als die Vermittlung beziehungsweise Aneignung der entsprechenden Grundlagen.

Gelegentlich entsteht hierbei der Eindruck, dass die Begriffe empirischer Forschung weitgehend unreflektiert übernommen werden. Betont werden muss jedoch: Nicht jede beliebige Frage ist eine Testaufgabe. Nicht jedes Auszählen von Ankreuzungen ist Kompetenzdiagnostik und damit ein belastbarer empirischer Beleg. Grundlegende Fragen der Stichprobenziehung, der Fragebogen- und Testkonstruktion und der Adäquatheit der gewählten statistischen Analysemethoden müssen einem breiteren Publikum zugänglich werden. So erkennen auch kritisch eingestellte Vertreter der Erziehungswissenschaft, dass insbesondere für den Bereich der Large-Scale-Assessments „ein erhebliches psychologisch-pädagogisches Methodenwissen in bestimmten statistischen Verfahren“ unabdingbare Voraussetzung ist (Merkens, 2006, S. 15). „Ein möglicher Beitrag der Erziehungswissenschaft oder Schulpädagogik ist […] bisher eher marginal geblieben, weil es auch an einer entsprechenden Ausbildung in den Studiengängen der Erziehungswissenschaft gemangelt hat“ (Merkens, 2006, S. 15). Auch Vertreter der empirischen Bildungsforschung selbst betonen die Schlüsselrolle der methodischen Ausbildung innerhalb der Nachwuchsförderung (vgl. Prenzel, 2006, S. 77).

Eine dritte – wenn auch sehr unwahrscheinliche – Zukunftsperspektive der Erziehungswissenschaft wäre die vollständige Loslösung vom traditionellen Selbstverständnis

und eine vollständige Überführung der Erziehungswissenschaft in die (empirische) Bildungsforschung. Tatsächlich plädieren einige Vertreter der Disziplin unter Bezugnahme auf den quantitativen wie qualitativen Einfluss der empirischen Bildungsforschung bereits heute für eine Umbenennung der Erziehungswissenschaft: „Wie es in den sechziger Jahren richtig war, im Zuge der realistischen Wende die einstmalige Pädagogik in Erziehungswissenschaft umzutaufen, so könnte es heute notwendig sein, einen neuen programmatisch aussagekräftigen Namen für eine zur Zeit offensichtlich falsch benannte Disziplin zu finden“ (Liebau, 2002, S. 296). Dieser Vorschlag zielt auf die Bezeichnung *Bildungswissenschaft* und wird von einigen Vertretern der Erziehungswissenschaft ernsthaft und unter Verweis auf weitere, inhaltlich relevante Argumente erwogen. Berücksichtigt man beispielsweise, dass gegenwärtig die Erwachsenenbildung und Weiterbildung den – an Teilnehmerzahlen und Finanzen gemessen – größten Bildungsbereich darstellt, so wird deutlich, dass *Erziehung* als Gegenstandsbestimmung an ihre Grenzen stößt, da man in diesem Kontext in der Regel von einem „erzogenen Menschen“ ausgehen muss, der Bildung und Lernen anstrebt, ohne dass hierbei Erziehungsaufgaben bestehen (vgl. hierzu Sailer, 2007). Inwieweit eine solche Sichtweise anschlussfähig ist, muss an dieser Stelle offen bleiben.

Abschließend möchte ich zu der Frage, welche Zukunftsperspektiven für die (Allgemeine) Erziehungswissenschaft in Abgrenzung zur oder in Anbindung an die (empirische) Bildungsforschung denkbar sind, Folgendes festhalten.

Wenn von empiriekritischen Vertretern der allgemeinen Erziehungswissenschaft gegenüber den aktuellen Tendenzen verstärkter Output-Orientierung, Standardisierung und hiermit einhergehend zunehmender Empirisierung hinsichtlich bildungsbezogener Fragen weiterhin undifferenziert auf Konfrontation und Abgrenzung beharrt wird, ohne sich auf eine konstruktive Auseinandersetzung mit diesen Strömungen einzulassen, wird dies meiner Ansicht nach zu einem weiteren Verlust an gesellschaftlicher und wissenschaftlicher Wertschätzung der Disziplin führen. Der bildungsbezogene gesellschaftliche und politische Informationsbedarf kann nur durch empirische Forschung befriedigt werden. Hierfür wäre es meines Erachtens nicht nur wünschenswert, sondern notwendig, die theoretischen Wissensbestände und praxisbezogenen Erfahrungen der Erziehungswissenschaft stärker in empirische Untersuchungen einfließen zu lassen. In diesem Sinne besteht die Hoffnung, die Interdisziplinarität der empirischen Bildungsforschung gleichberechtigt auszubauen und den erziehungswissenschaftlichen Anteil in der Bildungsforschung weiter zu stärken.

4.6 Kompetenzentwicklung

4.6.1 Das Problem der Modellierung von Kompetenzentwicklungstrends

Relevant, aber in der aktuellen Debatte bzgl. der Modellierung und Messung von Kompetenzen in der deutschsprachigen Diskussion nur unzureichend thematisiert, sind konzeptionelle und methodische Probleme bei der Modellierung von längerfristigen Trendverläufen auf Individual- wie Systemebene.

Es ist das erklärte Ziel des Bildungsmonitorings auf Grundlage der länderübergreifenden Bildungsstandards, *Veränderungen* in der Leistungsfähigkeit des Bildungssystems und somit in den schulischen Leistungen der Schülerinnen und Schüler zu untersuchen. Deshalb bedarf das Thema der Trendmodellierung größerer Aufmerksamkeit.

Während bei der Untersuchung von individuellen Lernverläufen im Normalfall deutliche Veränderungen der Kompetenzstände innerhalb weniger Jahre zu verzeichnen sind, fallen Veränderungen auf Systemebene deutlich geringer aus und erstrecken sich über wesentlich längere Zeiträume. So werden beispielsweise im Rahmen der *Long-Term Trend* (LTT)-Analysen des *National Assessment of Educational Progress (NAEP)*, dem nationalen Bildungsmonitoring der USA, für die Entwicklung der Lesekompetenz für den Zeitraum von 1984 bis 2004 Veränderungen innerhalb von Zwei- bis Vierjahreszeiträumen berichtet, die sich in Effektstärken von $d = 0.00$ bis maximal $d = 0.08$ ausdrücken lassen (Mazzeo & von Davier, 2008, S. 15f.). Diese Veränderungen bewegen sich also lediglich in der Größenordnung von Null bis drei Punkten auf der Berichtsmetrik, wobei Veränderungen in beide Richtungen gefunden wurden. Die ermittelten Veränderungen gehen somit kaum über das hinaus, was im Rahmen von Zufallsschwankungen erwartbar ist. Erklärend äußern Mazzeo und von Davier hierzu, dass Veränderungen auf der Berichtsmetrik von ein oder zwei Punkten und damit Effektstärken von $d = 0.06$ selten statisch bedeutsam waren. Erst Differenzen von mehr als vier Punkten, was einer Effektstärke von $d > 0.14$ entspricht, wurde durchgehend als statistisch signifikant ausgewiesen (Mazzeo & von Davier, 2008, S. 13f.).

Entgegen der verbreiteten Vorstellung muss es also keineswegs so sein, dass ein kontinuierliches Bildungsmonitoring einen kontinuierlichen, positiven Entwicklungstrend nachzeichnet. Vielmehr scheinen Stagnation oder auch Rückschritte in den ermittelten Leistungen realistisch. Angesichts solch geringer Veränderungen und dem zusätzlichen Aspekt, dass sich intendierte Wirkungen auf Systemebene erst nach vielen Jahren oder womöglich Jahrzehnten als positive Entwicklungsverläufe von Schülerleistungen manifestieren, wird die Relevanz einer methodisch einwandfreien Trendmodellierung erkennbar.

Das methodische Grundproblem von Trendmodellierungen besteht darin, dass das zur Verlinkung der Messzeitpunkte eingesetzte Instrumentarium über die Zeit hinweg keine oder zumindest nur geringstmögliche Änderungen erfahren sollte. Diesem Ansatz liegt der vielzitierte Ausspruch von Beaton: „When measuring change, do not change the measure" (Beaton, 1990, S. 165) zugrunde, welcher sich historisch wie folgt einordnen lässt:

Im Rahmen von NAEP, welches seit Anfang der 1970er Jahre regelmäßig Schülerleistungen in unterschiedlichen Jahrgängen und Inhaltsdomänen überprüft, wurden für das Jahr 1986 im Bereich des Leseverstehens unerwartete und wenig plausible Ergebnisse gefunden. Die so genannte *1986 NAEP Reading Anomaly* umschreibt das Phänomen, dass sich für 17jährige Schülerinnen und Schüler eine unplausible Verschlechterung der Leseleistungen von 1984 bis 1986 ergab, wobei sich gleichzeitig die Leseleistung der 13jährigen in unerklärlichem Umfang verbesserte, während für 9jährige nur unbedeutende Effekte gefunden wurden (vgl. Zwick, 1992). Die hier beobachteten Leistungsunterschiede belaufen sich auf Effekte von bis zu $d = 0.22$ und sind somit deutlich größer als alle üblicherweise in NAEP als Trend beobachteten Veränderungen. Langwierige, intensive Untersuchungen und Zusatzstudien ergaben, dass die Ursachen der Anomalie in verschiedenen Veränderungen des Testdesigns sowie der Testadministration begründet liegen, wobei von einem unentwirrbaren Konglomerat von Effekten auszugehen ist. Es werden unter anderem die folgenden Aspekte als ursächlich angenommen (vgl. Mazzeo & von Davier, 2008; Zwick, 1992):

- Obwohl Stimuli und Items identisch übernommen wurden, gab es Änderungen in der Zusammenstellung von Testblöcken und somit Reihenfolge- und Kontexteffekte.
- Die Zusammenstellung der Testdomänen wurde verändert: Während die Lesekompetenz 1984 gemeinsam mit der Schreibkompetenz der Schülerinnen und Schüler getestet wurde, waren 1986 neben den Leseitems Aufgaben zu Mathematik und Naturwissenschaften in den Testheften enthalten.
- Die für die Testblöcke angesetzten Lese- und Bearbeitungszeiten waren nicht identisch – allerdings beliefen sich die Unterschiede pro Block auf nur jeweils ein bis zwei Minuten. Da 1986 geringfügig mehr Zeit zur Verfügung stand, wurde zu diesem Messzeitpunkt auch die Zahl der pro Block zu bearbeitenden Items geringfügig erhöht.
- Das Layout der Lesetexte wurde verändert. Die Anpassungen betrafen die Druckfarbe, die Zeilenlänge und damit Zeilenumbrüche der Lesetexte sowie das Antwortformat geschlossener Items (korrekte Antwort einkreisen bzw. ankreuzen).

Eine wesentliche Quelle sich wandelnder Kontexteinflüsse ist ferner der Einsatz von *Mixed* im Gegensatz zu *Focused Designs*. In Mixed Designs werden pro Testheft Blöcke aus verschiedenen Kompetenzbereichen (bspw. Zuhören und Lesen) oder sogar aus verschiedenen Domänen (bspw. Naturwissenschaften, Mathematik und Lesen) administriert. Solche Mixed Designs kommen unter anderem in PISA und TIMSS zum Einsatz (Mazzeo & von Davier, 2008, S. 5). In einem Focused Design finden sich hingegen innerhalb eines Testhefts ausschließlich Blöcke eines Kompetenzbereichs aus einer einzigen Inhaltsdomäne. Ein solches Design wird zum Beispiel im Rahmen von Main NAEP und seit dem Jahr 2003 für die LTT-Analysen eingesetzt (Mazzeo & von Davier, 2008, S. 5).

Dass die Verwendung von Mixed Designs selbst in solchen Large-Scale-Assessments eine problematische Rolle spielt, die bereits durch umfangreiche Erfahrungen optimiert wurden und mit großer, internationaler Expertise durchgeführt werden, zeigt das Gutachten von Mazzeo und von Davier (2008), welche der PISA-Studie zwar Arbeit auf höchstem methodischem Niveau bescheinigen, aber dennoch Probleme für die Messung von Trends im PISA-Design identifizieren.

Auch im Rahmen der Pilotierung und Normierung der Bildungsstandards, in den zugehörigen Ländervergleichen sowie in den jährlich vom IQB durchgeführten Pilotierungen der Aufgaben für die Vergleichsarbeiten (VERA) in der dritten und achten Jahrgangsstufe im Fach Deutsch kommen wechselnde Mixed Designs zum Einsatz. Diese inkonstante kontextuelle Einbettung hat Einfluss auf die Schätzung der Itemparameter und somit auf die Bestimmung der Personenfähigkeiten.

Empfehlungen zur Behebung dieser Problematik der Testdesigns beziehen sich unter anderem auf den Einsatz eines Focused Designs, sowie auf die Verwendung einer großen Menge von Linkitems, wobei sich jeweils möglichst wenige Items auf denselben Stimulustext beziehen sollten (Mazzeo & von Davier, 2008)[3].

Ein Wechsel von Mixed zu Focused Designs bei der Evaluierung der Bildungsstandards sowie den zugehörigen Ländervergleichen ist allerdings nur schwer zu realisieren, da stets die Überprüfung mehrerer Testdomänen zu einem Messzeitpunkt angestrebt wird. Ferner ist für den deutschen Sprachraum aus Forschungsperspektive von Interesse, die Struktur der jeweils betrachteten Kompetenzen untersuchen zu können, was nur über die gemeinsame

[3] Die Beschränkung beziehungsweise Verringerung der Itemzahl pro Stimulustext würde dazu führen, dass notwendigerweise die Länge der Textpassagen verkürzt werden müsste, um weiterhin eine ökonomische Testung gewährleisten zu können. Dies könnte allerdings dazu führen, dass aus der Berücksichtigung ausschließlich kurzer Stimulustexte eine (unerwünschte) Veränderung des gemessenen Lesekompetenzkonstrukts resultiert. Diese Gefahr wird von den Autoren selbst erkannt und diskutiert (vgl. Mazzeo & von Davier, 2008, S. 28).

Administration der fraglichen Bereiche beziehungsweise Domänen bei denselben Personen durch die Generierung einer Kovarianzmatrix erfolgen kann. Ein möglicher Lösungsansatz könnte hier der Einsatz eines über Studien hinweg *konstanten Mixed Designs* sein, welches beispielsweise in einem Testheft immer die gemeinsame Administration von Blöcken aus denselben zwei Kompetenzbereichen gestattet. Dies könnte für das Fach Deutsch zum Beispiel bedeuteten, dass in einem Testheft mit vier Testblöcken stets zwei Blöcke zum Kompetenzbereich Lesen und zwei Blöcke zum Kompetenzbereich Zuhören an definierten Positionen zum Einsatz kommen. Auch in Ländervergleichen könnten dann stets mehrere Kompetenzbereiche evaluiert werden, dies allerdings in festgelegter und gleichbleibender Kombination. Die künftige Untersuchung der strukturellen Beziehungen der verschiedenen sprachlichen Kompetenzen könnte in separaten Studien zu Forschungszwecken und nicht im Rahmen des Monitorings des Bildungssystems erfolgen. Alternativ wären Designvarianten denkbar, die für eine Verlinkung verschiedener Studien sowohl ein konstantes Mixed Design als auch entsprechend der jeweiligen Forschungsfragen variable Designanteile umfasst.

Der für die LTT-Analysen in NAEP bis 1999 gewählte Zugang zur Lösung der oben angesprochenen methodischen Probleme durch die Veränderungen der Testinstrumente, lautet „[…] administering the exact same assessment booklets in successive assessment cycles using exactly the same sampling, administration, scoring, analysis, and quality monitoring procedures" (Mazzeo & von Davier, 2008, S. 15). Dies stellt in Anbetracht eines auftretenden Veränderungs*bedarfs* allerdings keine befriedigende Lösung dar, da die Notwendigkeit zur Veränderung der Messinstrumente in gewisser Weise in der Messung selbst angelegt ist, wie nachfolgend kurz skizziert werden soll.

Möchte man intraindividuelle Entwicklungsverläufe von Kompetenzständen nachzeichnen, so geht man von einer Veränderung und zwar einer Zunahme der zu messenden Kompetenzen über die Zeit aus. Dies bedeutet, dass sich auch das eingesetzte Messinstrument mit dem zunehmenden Kompetenzstand verändern müsste.

Ebenso möchte man den allmählichen unterrichtlichen Wandel hin zu einer zunehmenden Kompetenzorientierung auch in den Testinstrumenten abbilden. Dies könnte beispielsweise eine stärker integrativ angelegte Überprüfung von Kompetenzen im Bereich „Sprache und Sprachgebrauch untersuchen" betreffen. Auch um derartige Entwicklungen abbilden zu können, müssten Veränderungen der Messung über die Zeit hinweg erfolgen.

Ferner ergeben sich Fortschritte hinsichtlich der methodischen und testadministrativen Möglichkeiten, die durch ihre Nutzung ebenfalls eine Veränderung der Messmethoden mit sich bringen würden. Dies betrifft zum Beispiel den Einsatz computerbasierter Testmethoden.

Zwick (1992) umreißt das hier angedeutete Dilemma wie folgt: Es bleibt unklar „[…] how to measure performance change while remaining responsive to advances in curriculum and the technology of assessment" (Zwick, 1992, S. 206).

4.6.2 Methodische Defizite im Rahmen der Vergleichsarbeiten im Fach Deutsch in den Jahrgangsstufen 3 und 8

Veränderungen der Testinstrumente und Testadministration kommen im Zusammenhang mit den durch das IQB durchgeführten Testungen insbesondere im Kontext der jährlichen, flächendeckenden Vergleichsarbeiten (VERA) zum Tragen.

Hier findet zum Beispiel im Fach Deutsch sowohl in den Vergleichsarbeiten für die dritte (VERA-3) wie auch für die achte Jahrgangsstufe (VERA-8) ein jährlicher Wechsel der zweiten Testdomäne neben der Lesekompetenz statt. So kamen beispielsweise in VERA-3 im Fach Deutsch im Jahr 2011 Aufgaben aus den Kompetenzbereichen Lesen und Schreiben zum Einsatz, während im Jahr 2012 die Bereiche Lesen und Sprachgebrauch und im Jahr 2013 die Bereiche Lesen und Zuhören getestet werden sollen. Dies hat zur Folge, dass sich durch die wechselnde Zusammenstellung der Testdomänen die oben angesprochenen Reihenfolge- und Kontexteffekte niederschlagen können. Dieser Umstand wird dann zum Problem, wenn Itemparameter ohne Berücksichtigung dieser Veränderungen unreflektiert beibehalten werden, wie dies für die zur Verankerung eingesetzten Aufgaben aus der Normierung der Bildungsstandards für das Fach Deutsch im Primarbereich im Rahmen von VERA-3 regelmäßig der Fall ist.

Auch die in den VERA-Studien von den Ländern häufig praktizierte Anpassung und Neuformatierung der vom IQB pilotierten Stimuli sowie der dazugehörigen Items muss vor diesem Hintergrund nachdrücklich kritisiert werden. Es ist davon auszugehen, dass auch solche Veränderungen dazu führen, dass die in der Pilotierung ermittelten Itemparameter nur eingeschränkt auf die veränderten Aufgaben und Items übertragbar sind.

Zudem werden die Itemparameter für die in VERA eingesetzten Testaufgaben in einer Studie gewonnen, die eigentlich der Pilotierung der neu entwickelten Items und Aufgaben dient. Dadurch entspricht die Zusammenstellung der Testhefte in der eigentlichen VERA-Studie nicht der Testheftzusammenstellung, aus der die Itemparameter ermittelt werden. Dieses Vorgehen stellt ein schwerwiegendes testmethodisches Manko dar.

Empfehlenswert scheint hier allein ein Vorgehen, bei dem zunächst Stimuli und Items vom IQB in einer Pilotierung auf ihre grundsätzliche Verwendbarkeit hin geprüft und in einer nachfolgenden Normierungsstudie genau diejenigen Stimuli und Items in genau der Reihenfolge

und Zusammenstellung unter genau den Testbedingungen eingesetzt werden, die auch für die jeweilige Hauptstudie angedacht sind. Erst in einer so angelegten Normierung ließen sich Itemparameter gewinnen, die sowohl den Wünschen der Bildungsadministration als auch den methodischen Standards gerecht werden könnten.

Das im Rahmen der Vergleichsarbeiten aktuell gewählte Vorgehen entspricht im Lichte der oben geschilderten Befunde nicht dem gegenwärtigen methodischen Erkenntnisstand. Die mit den Vergleichsarbeiten intendierte Anbindung einer regelmäßigen Kompetenzdiagnostik an die jeweilige Bildungsstandardmetrik sowie die Verfolgung von Trendverläufen der schülerseitigen Kompetenzen erscheint somit fraglich.

Mazzeo und von Davier äußern hinsichtlich der in PISA und NAEP realisierten Testdesigns: „[…] we are cognisant of the fact that test designs and analysis procedures for all assessments are developed through trying to strike an appropriate balance among competing forces – the policy and informational goals of the assessments sponsors and participants, the practical and fiscal realities associated with actually carrying out the assessment, and the psychometric realities of what kinds of results can be reliably produced from data collected under a particular test design“ (Mazzeo & von Davier, 2008, S. 9). Damit bringen die Autoren auf den Punkt, was auch für die deutschen Vergleichsarbeiten in der dritten und achten Jahrgangsstufe zutrifft: Die Durchführung großer Schulleistungsstudien hat stets die Interessen und Perspektiven verschiedener Parteien zu vereinen. Obwohl der verantwortlichen wissenschaftlichen Institution also methodische Defizite in der gegenwärtigen Praxis der VERA-Studien bewusst sind, wird dennoch in der oben beschriebenen Weise verfahren. Dies erklärt sich unter anderem aus den inhaltlichen Vorgaben sowie der zeitlichen und finanziellen Budgetierung der VERA-Studien durch die Bildungsadministration.

Sollen die VERA-Studien ihren Zweck im Rahmen der Gesamtstrategie zur Qualitätsentwicklung erfüllen, so sollte (auch) von Seiten der Bildungsadministration erwogen werden, das Vorgehen in diesen Studien zu ändern. Möchte man daran festhalten, dass VERA jährlich durchgeführt wird, erscheint die zeitliche Ausdehnung der Entwicklungszyklen ein gangbarer Weg zu sein. Diese würde nicht nur für die Entwicklung von neuen, kompetenzorientierten und bildungsstandardkompatiblen Testaufgaben mehr Zeit einräumen, sondern würde ferner die Möglichkeit eröffnen, zunächst eine Pilotierung aller potentiell geeigneten Aufgaben und mit dem erforderlichen zeitlichen Abstand eine Normierung der final ausgewählten Aufgaben in genau der Testheftzusammenstellung durchzuführen, die auch für die eigentliche Hauptstudie geplant ist. Ein solches Vorgehen würde aber nicht nur eine zeitliche Ausweitung der Testentwicklungszyklen bedeuten, sondern auch zusätzliche finanzielle Mittel für die Durchführung der Normierungsstudien erfordern.

4.6.3 Die Entwicklung und Etablierung von Kompetenzmodellen und Kompetenzentwicklungsmodellen

Die Verabschiedung der länderübergreifenden Bildungsstandards sowie die Durchführung von stichprobenbasierten Ländervergleichen und regelmäßigen bildungsstandardbasierten Vergleichsarbeiten dienen letztlich dem Ziel, die Qualität der schulischen Bildung in Deutschland zu optimieren und auf einem möglichst hohen Niveau für alle Schülerinnen und Schüler dauerhaft zu stabilisieren. Die bei diesen Ansätzen leitende Output-Orientierung nimmt in den Blick, ob alle Kinder und Jugendlichen jene Kompetenzen erworben haben, die für den jeweiligen Zeitpunkt in ihrer Bildungsbiografie zu erwarten sind. Eine solche Perspektive ist eng mit dem Gedanken des kumulativen Kompetenzaufbaus verzahnt und erfordert regelmäßig verlässliche diagnostische Informationen zum aktuellen Leistungsstand der Schülerinnen und Schüler. Diese dient dem Ziel, Stärken und Schwächen identifizieren und durch gezielte Förderung am Ausbau jener Kompetenzen arbeiten zu können, die noch nicht dem angestrebten Niveau entsprechen.

Um eine solche Diagnostik entsprechend eines kumulativen Kompetenzerwerbs gewährleisten zu können, ist ein wichtiges Forschungsdesiderat der kommenden Jahre die Entwicklung und Etablierung von *Kompetenzentwicklungsmodellen*. Diese unterscheiden sich von Kompetenzmodellen und Kompetenzstufenmodellen, die gegenwärtig in der Deutschdidaktik sowie der empirischen Bildungsforschung diskutiert werden.

Kompetenzmodelle bilden die theoretische Fundierung der Auseinandersetzung mit einem Kompetenzbereich und beinhalten eine Konstruktbeschreibung, die den Ausgangspunkt für die Operationalisierung in Testinstrumenten darstellt. Ferner werden in einem Kompetenzmodell die verschiedenen Aspekte eines Kompetenzkonstrukts und die Zusammenhänge zwischen diesen Aspekten dargestellt.

Kompetenzstufenmodelle stellen eine Möglichkeit dar, das Kompetenzkontinuum in inhaltlich differenzierbare Bereiche zu unterteilen und somit querschnittlich ermittelte Leistungsstände so zu verorten, dass sie verständlich und anschaulich kommuniziert werden können. Auf diese Weise beschreiben sie aktuelle Unterschiede in den Leistungsständen von Schülerinnen und Schülern in einem breit definierten Kompetenzbereich. Es ist allerdings nicht davon auszugehen, dass Kompetenzstufenmodelle tatsächlich Entwicklungsstufen beschreiben und eine einheitliche Kompetenzentwicklung der Lernenden entlang der postulierten Stufen erfolgt. Bei der Rezeption von Kompetenzstufenmodellen ist zudem relevant, dass sich die Bestimmung der Stufengrenzen nie ausschließlich auf empirische Ergebnisse und fachdidaktische Erwägungen stützen kann, sondern stets auch normative und politische Vorgaben einschließt (vgl. Böhme & Köller, 2010).

Um eine Beschreibung des Erwerbs sprachlicher Kompetenzen über die Kinder- und Jugendphase bis zum frühen Erwachsenenalter im Sinne eines kumulativen Kompetenzerwerbs zu ermöglichen, sind *Kompetenzentwicklungsmodelle* erforderlich. Kompetenzdiagnostik, die sich auf ein Kompetenzentwicklungsmodell stützt, gestattet nicht nur zu erfassen, auf welche Kompetenzaspekte bereits zurückgegriffen werden kann, sondern ermöglicht zusätzlich die Feststellung, welches die nächsten Teilziele des Kompetenzaufbaus sein sollten. Damit bieten sie die Chance, auf den jeweiligen Entwicklungsstand von Schülerinnen und Schülern zugeschnittene Fördermaßnahmen einzuleiten. Eine solche unmittelbare Kopplung von Diagnostik und Förderung entspricht Wygotskis Konzept der *Zone der nächsten Entwicklung*, dem gemäß Bildungs- und Unterrichtsangebote auf das Entwicklungsniveau des Kindes abgestimmt sein müssen (vgl. Wygotski, 1987).

Kompetenzentwicklungsmodelle könnten somit eine wertvolle Perspektive für Diagnostik und Förderung darstellen. Hierfür bedürfen sie allerdings der engen interdisziplinären Kooperation, da eine fundierte theoretische Untermauerung durch die Deutschdidaktik ebenso unerlässlich ist wie eine methodisch einwandfreie empirische Erprobung. Um haltbare Kompetenzentwicklungsmodelle und entsprechende diagnostische Instrumente konstruieren zu können, müssen theoretische Durchdringung und empirische Überprüfung in einem iterativen Prozess ineinandergreifen.

Im Rahmen deutschdidaktischer Forschungsarbeit wurde insbesondere die Kompetenzentwicklung im Bereich Schreiben und hier sowohl die Entwicklung des freien Schreibens als auch der orthografischen Kompetenz detailliert beschrieben (vgl. für einen Überblick Böhme, Bremerich-Vos & Robitzsch, 2009; Böhme & Bremerich-Vos, 2009). Somit könnte der Bereich der Schreibkompetenz einen interessanten Ausgangspunkt für die Erstellung von Entwicklungsmodellen sprachlicher Kompetenzen darstellen.

Entscheidend ist allerdings, hierbei einen angemessenen Auflösungsgrad zu wählen. Das so genannte Bandbreiten-Genauigkeits-Dilemma (*Bandwidth-Fidelity-Dilemma;* Ones & Viswesvaran, 1996) umschreibt in der Diagnostik und Testkonstruktion das Prinzip, dass bei gegebenen Ressourcen, also beispielsweise einer beschränkten Testzeit, zwischen der sehr präzisen und reliablen Erfassung eines schmalen Fähigkeitsausschnitts und der weniger scharfen Erfassung eines breiteren Kompetenzspektrums abgewogen werden muss. Während nun aus diagnostischer Perspektive eher das Bedürfnis besteht, möglichst alle relevanten Fähigkeitsaspekte eines Konstrukts in eine Messung einzubeziehen, um das Konstrukt valide abbilden zu können (vgl. Abschnitt 2.4), tendieren Vertreter der Deutschdidaktik ebenso wie Lehrkräfte dazu – gleichzeitig – einen sehr hohen Auflösungsgrad der Messung anzustreben. Entsprechend betonen Abraham und Kollegen bezüglich der in den Bildungsstandards

ausgewiesenen Kompetenzbereiche: „Je nach dem Gegenstand des Unterrichts werden sie dabei unter dem Kriterium der Fachlichkeit sehr viel genauer zu fassen sein, als etwa die übergreifenden Kompetenzmodelle in der Bildungsforschung gegenwärtig sind. […] Nur unter dieser Voraussetzung können Kompetenzen wirklich differenziert erfasst werden“ (Abraham et al., 2007, S. 7). Nachfolgend weisen die Autoren darauf hin, dass solche ausdifferenzierten Kompetenzbeschreibungen empirisch abgesichert werden müssen. Testdiagnostisch ist es jedoch aufgrund von begrenzten Ressourcen, wie der verfügbaren Testzeit, zumeist nicht möglich, gleichzeitig sehr präzise und erschöpfend breit zu messen. Ferner zeigten die Analysen von Böhme und Bremerich-Vos (2009), dass auch theoretisch äußerst elaborierte Kompetenzmodelle einen Detailliertheitsgrad erreichen können, der in empirischen Daten nur schwer nachweisbar ist – zumindest mit den gegenwärtig verfügbaren Analysemethoden.

5 Literatur

Anmerkung: Die hier verzeichneten Literaturangaben betreffen ausschließlich die in der Rahmung der vorliegenden Arbeit zitierten Quellen. Die in den Einzelbeiträgen zitierten Quellen werden in jeweils separaten Literaturverzeichnissen im Anschluss an die Einzelbeiträge aufgeführt.

Abraham, U., Bremerich-Vos, A., Frederking, V. & Wieler, P. (2003). (Hrsg.). *Deutschdidaktik und Deutschunterricht nach PISA*. Freiburg im Breisgau: Fillibach Verlag.

Abraham, U., Baurmann, J., Feilke, H., Kammler, C. & Müller, A. (2007). Kompetenzorientiert unterrichten. Überlegungen zum Schreiben und Lesen. *Praxis Deutsch*, *203*, 6–14.

Abraham, U. & Müller, A. (2009). Aus Leistungsaufgaben lernen. *Praxis Deutsch*, *214,* 4-12.

van Ackeren, I. (2007). *Nutzung großflächiger Tests für die Schulentwicklung. Exemplarische Analyse der Erfahrungen aus England, Frankreich und den Niederlanden*. Berlin: Bundesministerium für Bildung und Forschung.

Anderson, J. R. (2001). *Kognitive Psychologie* (3. Aufl.). Heidelberg: Spektrum Akademischer Verlag.

Andresen, H. & Funke, R. (2003). Entwicklung sprachlichen Wissens und sprachlicher Bewusstheit. In U. Bredel, H. Günther, P. Klotz, J. Ossner & G. Siebert-Ott (Hrsg.), *Didaktik der deutschen Sprache. Ein Handbuch* (1. Teilband, S. 438-451). Paderborn: Schöningh.

Arbeitsgruppe „Internationale Vergleichsstudie“ (2007). *Vertiefender Vergleich der Schulsysteme ausgewählter PISA-Teilnehmerstaaten* (3. unveränderte Aufl.). Bonn: Bundesministerium für Bildung und Forschung (BMBF).

Artelt, C. & Schlagmüller, M. (2004). Der Umgang mit literarischen Texten als Teilkompetenz im Lesen? Dimensionsanalysen und Ländervergleiche. In U. Schiefele, C. Artelt, W. Schneider & P. Stanat (Hrsg), *Struktur, Entwicklung und Förderung von Lesekompetenz – Vertiefende Analysen im Rahmen von PISA 2000* (S. 169-196). Wiesbaden: VS Verlag für Sozialwissenschaften.

Artelt, C., Stanat, P., Schneider, W. & Schiefele, U. (2001). Lesekompetenz: Testkonzeption und Ergebnisse. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider et al. (Hrsg.), *PISA 2000 – Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 69-137). Opladen: Leske + Budrich.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice: designing and developing useful language tests*. Oxford: Oxford University Press.

Bartnitzky, H. (2008). Reihenweise Fehlurteile. *Die Grundschulzeitschrift*, *217*, 14-17.

Baumert, J. (2001). Vergleichende Leistungsmessung im Bildungsbereich. *Zeitschrift für Pädagogik* (*43. Beiheft*), 13-36.

Baumert, J., Artelt, C., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U. et al. (Hrsg.). (2002). *PISA 2000: Die Länder der Bundesrepublik Deutschland im Vergleich*. Opladen: Leske + Budrich.

Baumert, J., Bos, W. & Lehmann, R. (Hrsg.). (2000a). *TIMSS/III: Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn* (Bd. 1., Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit). Opladen: Leske und Budrich.

Baumert, J., Bos, W. & Lehmann, R. H. (Hrsg.). (2000b). *TIMSS/III: Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn* (Bd. 2., Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe). Opladen: Leske + Budrich.

Baumert, J., Brunner, M., Lüdtke, O., & Trautwein, U. (2007). Was messen internationale Schulleistungsstudien? Resultate kumulativer Wissenserwerbsprozesse. Eine Antwort auf Heiner Rindermann. *Psychologische Rundschau*, *58*, 118-128.

Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W. et al. (Hrsg.). (2001). *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen: Leske + Budrich.

Baumert, J., Lehmann, R. H., Lehrke, M., Schmitz, B., Clausen, M., Hosenfeld, I. et al. (1997). *TIMSS - Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich. Deskriptive Befunde*. Opladen: Leske und Budrich.

Baumert, J., Lüdtke, O., Trautwein, U., & Brunner, M. (2009). Large-scale student assessment studies measure the results of processes of knowledge acquisition: Evidence in support of the distinction between intelligence and student achievement. *Educational Research Review*, *4,* 165-176.

Beaton, A. E. (1990). Introduction. In A. E. Beaton & R. Zwick (Hrsg.), *The effect of changes in the National Assessment. Disentangling the NAEP 1985-86 Reading Anomaly* (pp. 1-13). Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.

De Beaugrande, R.-A. & Dressler, W. (1981). *Einführung in die Textlinguistik*. Tübingen: Niemeyer.

Beck, B. & Klieme, E. (Hrsg.). (2007). *Sprachliche Kompetenzen: Konzepte und Messung. DESI-Studie.* Weinheim: Beltz.

Behrens, U., Böhme, K. & Krelle, M. (2009). Zuhören – Operationalisierung und fachdidaktische Implikationen. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 357-375). Weinheim: Beltz.

Behrens, U. & Eichler, W. (2008). Sprachbewusstheit messen. In A. Bremerich-Vos, D. Granzer & O. Köller (Hrsg.), *Lernstandsbestimmung im Fach Deutsch. Gute Aufgaben für den Unterricht* (S. 186-195). Weinheim: Beltz.

Behrens, U. & Eriksson, B. (2009). Sprechen und Zuhören. In A. Bremerich-Vos, D. Granzer, U. Behrens & O. Köller (Hrsg.), *Bildungsstandards für die Grundschule: Deutsch konkret* (S. 43–74). Berlin: Cornelsen.

Belgrad, J., Eriksson, B., Pabst-Weinschenk, M. & Vogt, R. (2008). Die Evaluation von Mündlichkeit. Kompetenzen in den Bereichen Sprechen, Zuhören und Szenisch Spielen. In Symposium Deutschdidaktik e.V. (Hrsg.), *Sonderheft zum 16. Symposium Deutschdidaktik Kompetenzen im Deutschunterricht* [Sonderheft]. *Didaktik Deutsch*, *14,* 20-45.

Bereiter, C. & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Blum, W. (2006). Einführung. In W. Blum, C. Drüke-Noe, R. Hartung & O. Köller (Hrsg.), *Bildungsstandards Mathematik: konkret - Aufgabenbeispiele, Unterrichtsanregungen, Fortbildungsideen* (S. 14-32). Berlin: Cornelsen-Scriptor.

BMBF (2007) siehe Bundesministerium für Bildung und Forschung (2007).

BMFSFJ (2005) siehe Bundesministerium für Familie, Senioren, Frauen und Jugend (2005).

Böhme, K. (2006). Testen: ja – Den Unterricht verarmen: nein. *Grundschule*, *38 (5),* 8-10.

Böhme, K. & Bremerich-Vos, A. (2009). Diagnostik der Rechtschreibkompetenz in der Grundschule – Konstruktprüfung mittels Fehler- und Dimensionsanalysen. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 330-356). Weinheim: Beltz.

Böhme, K., Bremerich-Vos, A. & Robitzsch, A. (2009). Aspekte der Kodierung von Schreibaufgaben. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-

Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 290-329). Weinheim: Beltz.

Böhme, K. & Köller, O. (2010). Kompetenzstufenmodelle für den Mittleren Schulabschluss. Einleitung. In O. Köller, M. Knigge & B. Tesch (Hrsg.), *Sprachliche Kompetenzen im Ländervergleich* (S. 33-35). Münster: Waxmann.

Böhme, K., Neumann, D. & Schipolowski, S. (2010). Beschreibung der im Ländervergleich im Fach Deutsch untersuchten Kompetenzen. In O. Köller, M. Knigge & B. Tesch (Hrsg.), *Sprachliche Kompetenzen im Ländervergleich* (S. 19-25). Münster: Waxmann.

Böhme, K. & Robitzsch, A. (2009a). Methodische Aspekte der Erfassung der Lesekompetenz. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 250-289). Weinheim: Beltz.

Böhme, K. & Robitzsch, A. (2009b). Das Problem lokaler Abhängigkeiten von Items – Fehleranalysen in der Rechtschreibdiagnostik. Vortrag auf der 9. Tagung der Fachgruppe Methoden und Evaluation, Bielefeld, 10. – 12. September 2009.

Böhme, K., Robitzsch, A. & Busè, A.-K. (2010). Zur Abgrenzung des Hörverstehens gegenüber dem Leseverstehen mit Hilfe schwierigkeitsbestimmender Merkmale bei der Entwicklung von Testaufgaben. In V. Bernius & M. Imhof (Hrsg.), *Zuhörkompetenz in Unterricht und Schule. Beiträge aus Wissenschaft und Praxis.* Göttingen: Vandenhoeck & Ruprecht.

Böhnisch, M. (2008). Diskussionslinien innerhalb der Kompetenzdebatte. Ein Strukturierungsversuch. In Symposium Deutschdidaktik e.V. (Hrsg.), Sonderheft zum 16. Symposium Deutschdidaktik Kompetenzen im Deutschunterricht [Sonderheft]. *Didaktik Deutsch*, *14*, 5-19.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bollen, K. A. & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*, 305-314.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*, 425-440.

Borsboom, D. (2008). Latent variable theory. *Measurement, 6*, 25-53.

Bos, W. , Lankes E. M., Prenzel, M., Schwippert, K., Walther, G. & Valtin, R. (Hrsg.). (2003). *Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich.* Münster: Waxmann.

Bos, W., Hornberg, S., Arnold, K.-H., Faust, G., Fried, L., Lankes, E.-V. et al. (2007). *IGLU 2006 – Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich.* Münster: Waxmann.

Bremerich-Vos, A. (2002). Empirisches Arbeiten in der Deutschdidaktik. In C. Kammler & W. Knapp (Hrsg.), *Empirische Unterrichtsforschung und Deutschdidaktik* (S. 16-29). Baltmannsweiler: Schneider Verlag Hohengehren.

Bremerich-Vos, A. (2009). Die Bildungsstandards Deutsch. In A. Bremerich-Vos, D. Granzer, U. Behrens & O. Köller (Hrsg.), *Bildungsstandards für die Grundschule: Deutsch konkret* (S. 14–42). Berlin: Cornelsen.

Bremerich-Vos, A. & Böhme, K. (2009a). Lesekompetenzdiagnostik – die Entwicklung eines Kompetenzstufenmodells für den Bereich Lesen. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 250-289). Weinheim: Beltz.

Bremerich-Vos, A. & Böhme, K. (2009b). Kompetenzdiagnostik im Bereich „Sprache und Sprachgebrauch untersuchen". In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 376-392). Weinheim: Beltz.

Bremerich-Vos, A. & Wieler, P. (2003). Zur Einführung. In U. Abraham, A. Bremerich-Vos, V. Frederking & P. Wieler (Hrsg.), *Deutschdidaktik und Deutschunterricht nach PISA* (S. 13-25). Freiburg im Breisgau: Fillibach Verlag.

Bremerich-Vos, A., Böhme, K. & Robitzsch, A. (2009). Sprachliche Kompetenzen im Fach Deutsch – Strukturanalysen und Validierungsbefunde. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 198-218). Weinheim: Beltz.

Bremerich-Vos, A., Behrens, U., Böhme, K., Engelbert, M., Linkert, D. & Krelle, M. (2010). *Vergleichsarbeiten 2010. 3. Jahrgangsstufe (VERA-3). Deutsch – Didaktische Handreichung zu Testheft II – Rechtschreibung.* Berlin: Institut zur Qualitätsentwicklung im Bildungswesen (IQB).

Bremerich-Vos, A., Granzer, D., Behrens, U. & Köller, O. (Hrsg.). (2009). *Bildungsstandards für die Grundschule: Deutsch konkret.* Berlin: Cornelsen Scriptor.

Brezinka, W. (2004). Über die Krise der Pädagogik und ihre Zukunft als Universitätsfach. In P. Korte (Hrsg.), *Kontinuität, Krise und Zukunft der Bildung* (S. 29-40). Münster: Lit Verlag.

Britting, G. (1929). Brudermord im Altwasser. *Münchner Neueste Nachrichten*, (*Nr.317 vom 21.11.1929*)

Büker, P. & Vorst, C. (2010). Kompetenzen und Unterrichtsziele im Lese- und Literaturunterricht der Grundschule. In M. Kämper-van den Boogaart & K. H. Spinner (Hrsg.), *Lese- und Literaturunterricht* (Bd. 11/2, S. 21-48). Baltmannsweiler: Schneider Verlag.

Bundesministerium für Bildung und Forschung (Hrsg.). (2007). *Bildungsforschung Band 2: Vertiefender Vergleich der Schulsysteme ausgewählter PISA-Teilnehmerstaaten* (3. unveränderte Aufl.). Berlin: Bundesministerium für Bildung und Forschung.

Bundesministerium für Familie, Senioren, Frauen und Jugend (2005). *Zwölfter Kinder- und Jugendbericht. Bericht über die Lebenssituation junger Menschen und die Leistungen der Kinder- und Jugendhilfe in Deutschland.* Berlin: Bundesministerium für Familie, Senioren, Frauen und Jugend (BMFSFJ).

Busch, A. & Stenschke, O. (2008). *Germanistische Linguistik. Eine Einführung* (2., durchgesehene und korrigierte Aufl.). Tübingen: Narr.

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*, 1-47.

Cheng, L., Watanabe, Y. & Curtis, A. (Eds). (2004). *Washback in language testing: research contexts and methods.* Mahwah, N.J.: Lawrence Erlbaum and Associates.

Chomsky, N. (2006). *Language and mind.* Cambridge: University Press.

Christmann, U. (2010). Lesepsychologie. In M. Kämper-van den Boogaart & K. H. Spinner (Hrsg.), *Lese- und Literaturunterricht* (Bd. 11/1, S. 148-200). Baltmannsweiler: Schneider Verlag Hohengehren.

Christmann, U. & Groeben, N. (1997). Sprache. In J. Straub, W. Kempf & H. Werbik (Hrsg.), *Psychologie. Eine Einführung. Grundlagen, Methoden, Perspektiven* (S. 345-373). München: Deutscher Taschenbuch Verlag.

CollegeBoard (2011). Zugriff am 31.10.2011 unter http://www.collegeboard.org

Dedering, K. & Holtappels, H. G. (2010). Schulische Bildung. In R. Tippelt & B. Schmidt (Hrsg.), *Handbuch der Bildungsforschung* (3., durchgesehene Aufl., S. 365-382). Wiesbaden: Verlag für Sozialwissenschaften.

DESI-Konsortium (Hrsg.). (2008). *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie.* Weinheim: Beltz.

Deutscher Bildungsrat (1974): *Empfehlungen der Bildungskommission. Aspekte für die Planung der Bildungsforschung.* Verabschiedet auf der 37. Sitzung der Bildungskommission, 24./25./26.01.74 in Berlin. Bonn: Deutscher Bildungsrat.

Dürscheid, C. (2007). Damit das grammatische Abendland nicht untergeht. Grammatikunterricht auf der Sekundarstufe II. In K.-M. Köpcke & A. Ziegler (Hrsg.), *Grammatik in Universität und Schule. Theorie, Empirie und Modellbildung* (S. 45-65). Tübingen: Niemeyer.

EAA (2010). *ICAS Test and Subjects. English.* Zugriff am 31.10.2011 unter http://www.eaa.unsw.edu.au/about_icas/english

Eberl, M. (2004). Formative und reflektive Indikatoren im Forschungsprozess: Entscheidungsregeln und die Dominanz des reflektiven Modells. Schriften zur Empirischen Forschung und Quantitativen Unternehmensplanung der Ludwig-Maximilians-Universität München, *Heft 19/2004*.

Eckes, T. & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing, 23*, 290–325.

Ehlich, K. (1983). Text und sprachliches Handeln. Die Entstehung von Texten aus dem Bedürfnis nach Überlieferung. In A. Assmann, J. Assmann & C. Hardmeier (Hrsg.), *Schrift und Gedächtnis. Beiträge zur Archäologie der literarischen Kommunikation* (S. 24-43). München: Fink.

Ehlich, K. (2009). Sprechen im Deutschunterricht – didaktische Denkanstöße. In M. Krelle & C. Spiegel (Hrsg.), *Sprechen und Kommunizieren. Entwicklungsperspektiven, Diagnosemöglichkeiten und Lernszenarien in Deutschunterricht und Deutschdidaktik* (S. 8-14). Baltmannsweiler: Schneider Verlag Hohengehren.

Eichler, W. & Nold, G. (2007). Sprachbewusstheit. In B. Beck & E. Klieme (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie* (S. 63-82). Weinheim: Beltz.

Eichler, W. & Nold, G. (2006). Sprachbewusstheit. In E. Klieme & B. Beck (Hrsg.), *Sprachliche Kompetenzen – Konzepte und Messung* (S. 63-82). Weinheim: Beltz.

Embretson, S. E. & Reise, S. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum Publishers.

Emmrich, R. (2010). *Rückmeldungen VERA 8. Rückmeldeformate und Nutzungsmöglichkeiten. Schuljahr 2009/10.* (Vortrag auf einer Informationsveranstaltung zu Rückmeldungen im Rahmen von VERA 8/2010 des Instituts für Schulqualität der Länder Berlin und Brandenburg e.V. [ISQ]). Zugriff am 6.09.2011 unter http://www.isq-bb.de/uploads/media/VERA8_2010_Rueckmeldungen_Engl.pdf

EMSE-Netzwerk (2008). *Nutzung und Nutzen von Schulrückmeldungen im Rahmen standardisierter Lernstandserhebungen / Vergleichsarbeiten. Zweites Positionspapier des EMSE-Netzwerkes – verabschiedet auf der 9. EMSE-Fachtagung am 16. – 17. Dezember 2008 in Nürnberg.* Zugriff am 31.10.2011 unter http://vera-web.uni-landau.de/verapub/fileadmin/downloads/EMSE_Position2_Nutzung_VERA.pdf

Europarat (2001). *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen.* Berlin: Langenscheidt.

Felder, E. (2003). Sprache als Medium und Gegenstand des Unterrichts. In U. Bredel, H. Günther, P. Klotz, J. Ossner & G. Siebert-Ott (Hrsg.), *Didaktik der deutschen Sprache. Ein Handbuch* (1. Teilbd., S. 42-51). Paderborn: Schöningh.

Fiege, C., Reuther, F. & Nachtigall, C. (im Druck). Faire Vergleiche? - Berücksichtigung von Kontextbedingungen des Lernens beim Vergleich von Testergebnissen aus deutschen Vergleichsarbeiten. *Zeitschrift für Bildungsforschung.*

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37,* 359–374.

Flower, L. & Hayes J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, *32,* 365-387.

Frederking, V. (2008). Literarische bzw. (literar-)ästhetische Kompetenz. Möglichkeiten und Probleme der empirischen Erhebung eines Kernbereichs des Deutschunterrichts. In V. Frederking (Hrsg.), *Schwer messbare Kompetenzen. Herausforderungen für die empirische Fachdidaktik* (S. 36-64). Baltmannsweiler: Schneider Hohengehren.

Frederking, V. (2010). Modellierung literarischer Rezeptionskompetenz. In M. Kämper-van den Boogaart & K. Spinner (Hrsg.), *Lese- und Literaturunterricht* (Bd. 11/1, S. 324-380). Baltmannsweiler: Schneider Verlag Hohengehren.

Frederking, V., Hu, A., Krejci, M., Legutke, M., Oomen-Welke, I. & Vollmer, J. (2003). Thesen zum Sprachenunterricht und zu den Sprachdidaktiken. In H. Bayrhuber, B. Ralle, K. Reiss, L.-H. Schön & H.-J. Vollmer (Hrsg.), *Konsequenzen aus PISA. Perspektiven der Fachdidaktiken* (S. 75-82). Innsbruck: Studien Verlag.

Frederking, V., Meier, C., Roick, T., Steinhauer, L., Stanat, P. & Dickhäuser, O. (2009). Literarästhetische Urteilskompetenz erfassen. In A. Bertschi-Kaufmann & C. Rosebrock (Hrsg.), *Literalität: Bildungsaufgabe und Forschungsfeld* (S. 165-180). Weinheim: Juventa.

Friederici, A. D. (1984). *Neuropsychologie der Sprache.* Stuttgart: Kohlhammer.

Frey, A. (2007). Adaptives Testen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 261-278). Berlin: Springer.

Gailberger, S. (2008). Leseförderung durch Hörbücher. Eine verbal-auditive Leseförderungstheorie für den Deutschunterricht. In B. Lecke (Hrsg.), *Mediengeschichte, Intermedialität und Literaturdidaktik* (S. 395-446). Frankfurt am Main: Peter Lang.

Gailberger, S. (2010). Hörbücher und das simultane Lesen und Hören im Deutschunterricht. Befunde zu einer mehrdimensionalen Förderung von literarischen und Lesekompetenzen schwacher Schüler an der Schnittstelle von Schriftlichkeit und Mündlichkeit. In V. Bernius & M. Imhof (Hrsg.), *Zuhörkompetenz in Unterricht und Schule. Beiträge aus Wissenschaft und Praxis* (S. 105-134). Göttingen: Vandenhoeck & Ruprecht.

Ginther, A. & Stevens, J. (1998). Language background, ethnicity, and the internal construct validity of the Advanced Placement Spanish Language Examination. In A. J. Kunnan (Ed.), *Validation in language assessment: selected papers from the 17th Language Testing Research Colloquium Long Beach* (pp. 169–194). Mahwah, NJ: Lawrence Erlbaum.

Gloger-Tippelt, G. (2010). Kindheit und Bildung. In R. Tippelt & B. Schmidt (Hrsg.), *Handbuch der Bildungsforschung* (3., durchgesehene Aufl.) (S. 627-640). Wiesbaden: Verlag für Sozialwissenschaften.

Gornik, H. & Granzow-Emden, M. (2008). Sprachthematisierung und grammatische Begriffe. In Symposium Deutschdidaktik e.V. (Hrsg.), *Sonderheft zum 16. Symposium Deutschdidaktik Kompetenzen im Deutschunterricht* [Sonderheft]. *Didaktik Deutsch*, *14*, 127-138.

Götz, T., Frenzel, A. C. & Pekrun, R. (2010). Psychologische Bildungsforschung. In R. Tippelt & B. Schmidt (Hrsg.), *Handbuch der Bildungsforschung* (3., durchgesehene Aufl., S. 71-91). Wiesbaden: Verlag für Sozialwissenschaften.

Granzer, D., Böhme, K. & Köller, O. (2008). Kompetenzmodelle und Aufgabenentwicklung für die standardisierte Leistungsmessung im Fach Deutsch. In A. Bremerich-Vos, D. Granzer, & O. Köller (Hrsg.), *Lernstandsbestimmung im Fach Deutsch* (S. 10-28). Weinheim: Beltz.

Green, J. P., Winters, M. A. & Forster, G. (2003). *Testing high stakes tests: can we believe the results of accountability tests?* (Civic Report No. 33). New York: Manhattan Institute for Policiy Research.

Grimm, H. & Wilde, S. (1998). Sprachentwicklung: Im Zentrum steht das Wort. In H. Keller (Hrsg.), *Lehrbuch Entwicklungspsychologie* (S. 445-473). Bern: Huber.

Groeben, N. (2005). Auf dem Weg zu einer deutsch-didaktischen Unterrichtsforschung? In J. Stückrath & R. Strobel (Hrsg.), *Deutschunterricht empirisch. Beiträge zur Überprüfbarkeit von*

Lernfortschritten im Sprach-, Literatur- und Medienunterricht (S. 7-33). Baltmannsweiler: Schneider Verlag Hohengehren.

Groeben, N. & Hurrelmann, B. (Hrsg.). (2002). *Lesekompetenz. Bedingungen, Dimensionen, Funktionen*. Weinheim: Juventa.

Grotjahn, R. (2002). Konstruktion und Einsatz von C-Tests: Ein Leitfaden für die Praxis. In R. Grotjahn (Hrsg.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Bd. 4, S. 211-225). Bochum: AKS-Verlag.

Gruschka, A. (2006a): Bildungsstandards oder das Versprechen, Bildungstheorie in empirischer Bildungsforschung aufzuheben. *Pädagogische Korrespondenz*, (*35*), 5-22.

Gruschka, A. (2006b). Pädagogische Theorie und empirische Forschung. In G. Wiesner, C. Zeuner & H. J. Forneck (Hrsg.), *Empirische Forschung und Theoriebildung in der Erwachsenenbildung* (S. 34-48). Baltmannsweiler: Schneider Verlag Hohengehren.

Gruschka, A. (2007). Bildungsstandards oder das Versprechen, Bildungstheorie in empirischer Bildungsforschung aufzuheben. In L. A. Pongratz, R. Reichenbach, M. Wimmer (Hrsg.), *Bildung – Wissen – Kompetenz* (S. 9-29). Bielefeld: Janus Presse.

Grzesik, J. (2003). Was testet der PISA-Test des Lesens? In U. Abraham, A. Bremerich-Vos, V. Frederking & P. Wieler (Hrsg.), *Deutschdidaktik und Deutschunterricht nach PISA* (S. 135-164). Freiburg im Breisgau: Fillibach Verlag.

Grzesik, J. (2005). *Texte verstehen lernen*. Münster: Waxmann.

Habermas, J. (1974/1989). Notizen zur Entwicklung der Interaktionskompetenz. In J. Habermas, *Vorstudien und Ergänzungen zur Theorie des kommunikativen Handelns* (S. 187-225). Frankfurt/Main: Suhrkamp.

Habermas, J. (1981). *Theorie der kommunikativen Kompetenz*. Frankfurt/Main: Suhrkamp.

Habermas, J. (1982/1989). Erläuterungen zum Begriff des kommunikativen Handelns. In J. Habermas, *Vorstudien und Ergänzungen zur Theorie des kommunikativen Handelns* (S. 571-605). Frankfurt/Main: Suhrkamp.

Hansel, T. (2007). Zur Situation der Erziehungswissenschaft zwischen Legitimationskrise und Innovationsdynamik. In F.-M. Konrad & M. Sailer (Hrsg.), *Homo educabilis* (S. 175-198). Münster: Waxmann.

Hasselgren, A. (2002). Learner Corpora and language testing – Smallwords as markers of learner fluency. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 143-173). Amsterdam: John Benjamins.

Hattie, J., Biggs, J. & Purdie, N. (1996). Effects of learning skills interventions on student learning: A meta-analysis. *Review of Educational Research*, *66,* 99-136.

Hayes, J. R. & Flower, L. (1980). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3-30). Hillsdale, NJ: Lawrence Erlbaum Associates.

Heid, H. (1996). Über Zweifel an der Möglichkeit, Pädagogik als empirische Wissenschaft zu betreiben. In Max-Planck-Institut für Bildungsforschung (Hrsg.), *Pädagogik als empirische Wissenschaft. Reden zur Emeritierung von Peter Martin Roeder* (S. 17-60). Berlin: Max-Planck-Institut für Bildungsforschung.

Helmke, A., Helmke, T., Heyne, N., Hosenfeld, A., Kleinbub, I., Schrader, F.-W. & Wagner, W. (2007). Erfassung, Bewertung und Verbesserung des Grundschulunterrichts: Forschungsstand, Probleme und Perspektiven. In K. Möller, P. Hanke, C. Beinbrech, A. K. Hein, T. Kleickmann & R. Schages (Hrsg.), *Qualität von Grundschulunterricht* (S. 17-34). Wiesbaden: Verlag für Sozialwissenschaften.

Helmke, A. & Hosenfeld, I. (2005). Standardbezogene Unterrichtsevaluation. In G. Brägger, B. Bucher & N. Landwehr (Hrsg.), *Schlüsselfragen zur externen Schulevaluation* (S. 127-151). Bern: h.e.p.-Verlag.

Herrmann, T. (1992). Sprechen und Sprachverstehen. In H. Spada (Hrsg.), *Lehrbuch allgemeine Psychologie* (2., korrigierte Aufl., S. 281-322). Bern: Huber.

Herné, K.-L. & Naumann, C.L. (2005). *Aachener Förderdiagnostische Rechtschreibfehler-Analyse (AFRA). Systematische Einführung in die Praxis der Fehleranalyse mit Auswertungshilfen zu insgesamt 33 standardisierten Testverfahren als Kopiervorlagen* (4. Auflage). Aachen: Alfa Zentaurus.

Hornbostel, S. & Keiner, E. (2002). Evaluation der Erziehungswissenschaft. *Zeitschrift für Erziehungswissenschaft , 5*, 634-653.

Hosenfeld, I. & Groß Ophoff, J. (2007).Editorial. In I. Hosenfeld & J. Groß Ophoff (Hrsg.), *Nutzung und Nutzen von Evaluationsstudien in Schule und Unterricht* [Themenheft 4/2007]. *Empirische Pädagogik*, *21*, 352-367.

Hoskens, M. & De Boeck, P. (1995). Componential IRT models for polytomous items. *Journal of Educational Measurement*, *32*, 364-384.

Hoskens, M. & De Boeck, P. (1997). A parametric model for local dependence among test items. *Psychological Methods*, *2*, 261-277.

Hoskens, M. & De Boeck, P. (2001). Multidimensional componential item response theory models for polytomous items. *Applied Psychological Measurement*, *25*, 19-37.

Hurrelmann, B. (2007). Modelle und Merkmale der Lesekompetenz. In A. Bertschi-Kaufmann (Hrsg.), *Lesekompetenz – Leseleistung – Leseförderung* (S. 18-28). Seelze-Velber: Kallmeyer.

Hymes, D. H. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics: selected readings* (pp. 269-293). Harmondsworth: Penguin.

Imhof, M. (2003). *Zuhören. Psychologische Aspekte auditiver Informationsverarbeitung*. Göttingen: Vandenhoeck & Ruprecht.

Imhof, M. (2010). Zuhören lernen und lehren. Psychologische Grundlagen zur Beschreibung und Förderung von Zuhörkompetenzen in Schule und Unterricht. In V. Bernius & M. Imhof (Hrsg.), *Zuhörkompetenz in Unterricht und Schule. Beiträge aus Wissenschaft und Praxis* (S. 15-30). Göttingen: Vandenhoeck & Ruprecht.

Imhof, M. & Bernius, V. (2010). Zuhörkompetenz in Schule und Unterricht – Grundlagen und Erfahrungen. In V. Bernius & M. Imhof (Hrsg.), *Zuhörkompetenz in Unterricht und Schule. Beiträge aus Wissenschaft und Praxis* (S. 7-14). Göttingen: Vandenhoeck & Ruprecht.

Ingenkamp, K. & Lissmann, U. (2008). *Lehrbuch der Pädagogischen Diagnostik*. Weinheim: Beltz.

Isaac, K., Eichler, W. & Hosenfeld, I. (2008). Ein Modell zur Vorhersage von Aufgabenschwierigkeiten im Kompetenzbereich Sprache und Sprachgebrauch untersuchen. In B. Hofmann & R. Valtin (Hrsg.), *Checkpoint Literacy - Tagungsband 2 zum 15. Europäischen Lesekongress 2007 in Berlin* (S. 12-27). Berlin: Deutsche Gesellschaft für Lesen und Schreiben.

Jäger, R. S. (2003). Pädagogisch-psychologische Diagnostik. In K. D. Kubinger & R. S. Jäger (Hrsg.), *Schlüsselbegriffe der Psychologischen Diagnostik* (S. 313-316). Weinheim: Beltz.

Jäger, R. S., Frey, A., Wosnitza, M. & Flor, D. (2001). Pädagogische Diagnostik. In L. Roth (Hrsg.), *Pädagogik. Handbuch für Studium und Praxis* (S. 848–872). München: Oldenburg.

Jahnke, T. (2008). *Die empirische Wünschelrutengängerei und ihre Folgen*. Vortrag auf der 42. Tagung für Didaktik der Mathematik. Jahrestagung der Gesellschaft für Didaktik der Mathematik, 13.-18.3.2008, Budapest, Ungarn. Zugriff am 5.09.2011 unter http://www.mathematik.tu-dortmund.de/ieem/cms/media/BzMU/BzMU2008/BzMU2008/BzMU2008_JAHNKE_Thomas_CD.pdf

Jendrowiak, H.-W. (1998). Die Theorie der humanen Schule als pädagogische Theorie. In H.-W. Jendrowiak (Hrsg.), *Humane Schule in Theorie und Praxis* (S. 94-113). Frankfurt/Main: Peter Lang.

Jendrowiak, H.-W. (2001). Die Sinnfrage in der Pädagogik. Eine Einführung in die Thematik. *Vierteljahrsschrift für Wissenschaftliche Pädagogik*, *77*, 1-5.

Jude, N. (2008). *Zur Struktur von Sprachkompetenz.* Dissertation, Goethe-Universität Frankfurt a.M. Zugriff am 23.08.2011 unter: http://publikationen.ub.uni-frankfurt.de/volltexte/2009/6794/pdf/Jude_Zur_Struktur_von_Sprachkompetenz.pdf

Jude, N., Klieme, E., Eichler, W., Lehmann, R. H., Nold, G., Schröder, K. et al. (2008). Strukturen sprachlicher Kompetenzen. In DESI-Konsortium (Hrsg.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* (S. 191-201). Weinheim: Beltz.

Kahl, R. & Spiewak, M. (2005). Nur bedingt wissenschaftlich. Die Erziehungswissenschaften haben in der Forschung und der Lehrerausbildung versagt. Eine Polemik. *Die Zeit,* (*Nr. 11 vom 10.03.2005*). Zugriff am 18.08.2011 unter: http://www.zeit.de/2005/11/B-Erziehungswissenschaften

Kammler, C. & Knapp, W. (2002). Empirische Unterrichtsforschung als Aufgabe der Deutschdidaktik. In C. Kammler & W. Knapp (Hrsg.), *Empirische Unterrichtsforschung und Deutschdidaktik* (S. 2-14). Baltmannsweiler: Schneider Verlag Hohengehren.

Kämper-van den Boogaart, M. (2003). Lesekompetenzen – Hauptsache flexibel. Zu einer Parallele zwischen Literaturdidaktik und empirischer Lesepsychologie. In U. Abraham, A. Bremerich-Vos, V. Frederking & P. Wieler (Hrsg.), *Deutschdidaktik und Deutschunterricht nach PISA* (S. 26-46). Freiburg im Breisgau: Fillibach.

Kämper-van den Bogaart, M. & Pieper, I. (2008). Literarisches Lesen. In Symposium Deutschdidaktik e.V. (Hrsg.), *Sonderheft zum 16. Symposium Deutschdidaktik Kompetenzen im Deutschunterricht* [Sonderheft]. *Didaktik Deutsch, 14*, 46-65.

Karmiloff-Smith, A. (1992). *Beyond modularity: A developmental perspective on cognitive science.* Cambridge, Mass.: MIT Press/Bradford Books.

Kellogg, R. T. (1996). A model of working memory in writing. In C. M. Levy & S. E. Ransdell (Eds.), *The science of writing: theories, methods, individual differences, and applications* (pp. 57-71). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* New York: Cambridge University Press.

Kirsch, I. S., Jungeblut, A. & Mosenthal, P. B. (1998). The measurement of adult literacy. In T. S. Murray, I. S. Kirsch & L. B. Jenkins (Hrsg.), *Adult literacy in OECD countries* (pp. 105-134). Washington D.C.: U.S. Department of Education.

Klauer, K. J. (1982). *Handbuch der Pädagogischen Diagnostik* (Bd. 1 und 2). Düsseldorf: Schwann.

Klieme, E. (2007). Empirische Schulforschung versus Allgemeine Erziehungswissenschaft? Eine Erwiderung zum Statement von Jörg Ruhloff. In M. Kraul (Hrsg.), *Bildungsforschung und Bildungsreform* [Beiheft]. *Die Deutsche Schule*, *99,* 141-145.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M. et al. (2007). *Zur Entwicklung nationaler Bildungsstandards. Expertise.* Bonn: Bundesministerium für Bildung und Forschung (BMBF).

Klieme, E. & Hartig, J. (2007). Kompetenzkonzepte in den Sozialwissenschaften und im erziehungswissenschaftlichen Diskurs. In M. Prenzel, I. Gogolin & H.-H. Krüger (Hrsg.),*Kompetenzdiagnostik* [Sonderheft 8/2007]. *Zeitschrift für Erziehungswissenschaft*, *10*, 11-29.

Klieme, E., Hartig, J. & Rauch, D. (2008). The concept of competence in educational contexts. In J. Hartig, E. Klieme & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 3-22). Cambridge, Mass.: Hogrefe.

Klieme, E. & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingerichteten Schwerpunktprogramms der DFG. *Zeitschrift für Pädagogik*, *52*, 876-903.

Klieme, E. und Rakoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik. Outcome-orientierte Messung und Prozessqualität des Unterrichts. *Zeitschrift für Pädagogik*, *54*, 222-237.

Klieme, E., Stanat, P. & Artelt, C. (2001). Fächerübergreifende Kompetenzen. Konzepte und Indikatoren. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 203-218). Weinheim: Beltz.

KMK (1989). siehe Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (1989).

KMK (2004). siehe Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2004).

KMK (2005a). siehe Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2005a).

KMK (2005b). siehe Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2005b).

KMK (2006). siehe Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2006).

Koch, P. & Österreicher, W. (1985). Sprache der Nähe – Sprache der Distanz – Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanisches Jahrbuch*, *36*, 15-43.

Koch, U., Groß Ophoff, J., Hosenfeld, I. & Helmke, A. (2006). Von der Evaluation zur Schul- und Unterrichtsentwicklung - Ergebnisse der Lehrerbefragungen zur Auseinandersetzung mit den VERA-Rückmeldungen. In F. Eder, A. Gastager & F. Hofmann (Hrsg.), *Qualität durch Standards? Beiträge zur 67. Tagung der Arbeitsgruppe der Empirischen Bildungsforschung (AEPF), Salzburg* (S. 187-199). Münster: Waxmann.

Koeppen, K., Hartig, J., Klieme, E. & Leutner, D. (2008). Current issues in competence modeling and assessment. *Zeitschrift für Psychologie/Journal of Psychology, 216*, 61-73.

Köller, O. (2008). Bildungsstandards in Deutschland: Implikationen für die Qualitätssicherung und Unterrichtsqualität. In M. Meyer, M. Prenzel & S. Hellekamps (Hrsg.), *Perspektiven der Didaktik* [Sonderheft 9/2008]. *Zeitschrift für Erziehungswissenschaft, 11*, 47-59.

Köller, O. (2010). Bildungsstandards. In R. Tippelt & B. Schmidt (Hrsg.), *Handbuch der Bildungsforschung* (3., durchgesehene Aufl.) (S. 529-548). Wiesbaden: Verlag für Sozialwissenschaften.

Köller, O. & Trautwein, U. (2004). Englischleistungen von Schülerinnen und Schülern an allgemein bildenden und beruflichen Gymnasien. In O. Köller, R. Watermann, U. Trautwein & O. Lüdtke (Hrsg.), *Wege zur Hochschulreife in Baden-Württemberg. TOSCA - Eine Untersuchung an allgemein bildenden und beruflichen Gymnasien* (S. 285-326). Opladen: Leske + Budrich.

Köller, O., Watermann, R., Trautwein, U. & Lüdtke, O. (2004). *Wege zur Hochschulreife in Baden-Württemberg. TOSCA - Eine Untersuchung an allgemein bildenden und beruflichen Gymnasien*. Opladen: Leske + Budrich.

König, E. & Zedler, P. (2004). Erziehungswissenschaftliche Forschung in Deutschland. *Pädagogische Rundschau*, *58*, 75-91.

Koretz, D. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *Journal of Human Resources*, *37*, 752-777.

Koretz, D. (2005). Alignment, high stakes, and the inflation of test scores. *Yearbook of the National Society for the Study of Education*, *104*, 99-118.

Köster, J. (2008). Lernaufgaben - Leistungsaufgaben. *Deutschunterricht*, *61*, 4-11.

Köster, J. (2010). Aufgabentypen für Erfolgskontrollen und Leistungsmessung im Literaturunterricht. In M. Kämper-van den Boogaart & K. H. Spinner (Hrsg.), *Lese- und Literraturunterricht* (Bd. 11/3, S. 3-26). Baltmannsweiler: Schneider Verlag Hohengehren.

Köster, J. & Lindauer, T. (2008). Zum Stand wissenschaftlicher Aufgabenreflexion aus deutschdidaktischer Perspektive. In Symposium Deutschdidaktik e.V. (Hrsg.), *Sonderheft zum 16. Symposium Deutschdidaktik Kompetenzen im Deutschunterricht* [Sonderheft]. *Didaktik Deutsch*, *14*, 148-161.

Krelle, M. (2010). Zuhördidaktik. Anmerkungen zur Förderung rezeptiver Fähigkeiten des mündlichen Sprachgebrauchs im Deutschunterricht. In V. Bernius & M. Imhof (Hrsg.), *Zuhörkompetenz in Unterricht und Schule. Beiträge aus Wissenschaft und Praxis* (S. 51-68). Göttingen: Vandenhoeck & Ruprecht.

Kühn, P. (2008). „Gute Aufgaben" zur Lernstandsbestimmung im Kompetenzbereich „Sprache und Sprachgebrauch untersuchen". In A. Bremerich-Vos, D. Granzer & O. Köller (Hrsg.), *Lernstandsbestimmung im Fach Deutsch. Gute Aufgaben für den Unterricht* (S. 196-212). Weinheim: Beltz.

Kürschner, C. & Schnotz, W. (2008). Das Verhältnis gesprochener und geschriebener Sprache bei der Konstruktion mentaler Repräsentationen. *Psychologische Rundschau*, *59*, 139–149.

Kyllonen, P. C., Walters, A. M. & Kaufman, J. C. (2011). *The role of noncognitive constructs and other background variables in graduate education.* Princeton, NJ.: Educational Testing Service.

Langfeldt, H. P. & Tent, L. (1999). *Pädagogisch-psychologische Diagnostik* (Bd. 2, Anwendungsbereiche und Praxisfelder). Göttingen: Hogrefe.

Lehmann, R. H., Peek, R. & Poerschke, J. (2006). *HAMLET 3–4. Hamburger Lesetest für 3. und 4. Klassen.* Göttingen: Hogrefe.

Leubner, M. (2005). Die neuen Bildungsstandards und die aktuellen Aufgaben in Deutschbüchern. In J. Stückrath & R. Strobel (Hrsg.), *Deutschunterricht empirisch* (S. 162-176.) Baltmannsweiler: Schneider Verlag Hohengehren.

Leucht, M., Retelsdorf, J., Möller, J. & Köller, O. (2010). Zur Dimensionalität rezeptiver englischsprachiger Kompetenzen. *Zeitschrift für Pädagogische Psychologie*, *24,* 123-138.

Leutner, D. (2001). Pädagogisch-psychologische Diagnostik. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (S. 521-530). Weinheim: PVU.

Liebau, E. (2002). Bildungswissenschaft. Zur Weiterentwicklung der Disziplin. *Vierteljahresschrift für wissenschaftliche Pädagogik*, *78*, 293-299.

Lipowsky, F. (2007). Unterrichtsqualität in der Grundschule – Ansätze und Befunde der nationalen und internationalen Forschung. In K. Möller, P. Hanke, C. Beinbrech, A. K. Hein, T. Kleickmann & R. Schages (Hrsg.), *Qualität von Grundschulunterricht* (S. 35-49). Wiesbaden: Verlag für Sozialwissenschaften.

Lorenz, C. & Artelt, C. (2009). Fachspezifität und Stabilität diagnostischer Kompetenz von Grundschullehrkräften in den Fächern Deutsch und Mathematik. *Zeitschrift für Pädagogische Psychologie*, *23*, 211-222.

Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.

Mazzeo, J., & von Davier, M. (2008). *Review of the Programme for International Student Assessment (PISA) test design: recommendations for fostering stability in assessment results.* Zugriff am 26.08.2011 unter http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=EDU/PISA/GB%282008%2928&docLanguage=En

McClelland, D. C. (1973). Testing for competencies rather than for intelligence. *American Psychologist*, *28*, 1-14.

Merkens, H. (2005). Unbedingt wissenschaftlich. *Die Zeit* (Nr. 13 vom 23.03.2005). Zugriff am 18.08.2011 unter http://www.zeit.de/2005/13/Replik

Merkens, H. (2006). Bildungsforschung und Erziehungswissenschaft. In H. Merkens (Hrsg.), *Erziehungswissenschaft und Bildungsforschung* (S. 9-20). Wiesbaden: Verlag für Sozialwissenschaften.

Messick, S. (1989a). Validity. In R. Linn (Ed.), *Educational measurement* (pp. 13-103). New York: Macmillan.

Messick, S. (1989b). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*, 5-11.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23,* 13-23.

NAPLAN (2010). *NAPLAN summary report. Achievement in reading, writing, language conventions and numeracy.* Zugriff am 23.08.2011 unter http://www.naplan.edu.au/verve/_resources/NAPLAN_2010_Summary_Report.pdf

National Assessment Governing Board (2008). *Reading framework for the 2009 National Assessment of Educational Progress.* Washington, DC: National Assessment Governing Board.

Naumann, C. L. (2008). Zur Rechtschreibkompetenz und ihrer Entwicklung. In A. Bremerich-Vos, D. Granzer & O. Köller (Hrsg.), *Lernstandsbestimmung im Fach Deutsch. Gute Aufgaben für den Unterricht* (S. 134-160). Weinheim: Beltz.

Nauwerck, P. (2009). Sprachstandsdiagnose und Sprachförderung im Übergang vom Kindergarten in die Grundschule: Förderbedarf erkennen und (kommunikative) Kompetenzen entwickeln. In M. Krelle & C. Spiegel (Hrsg.), *Sprechen und Kommunizieren. Entwicklungsperspektiven, Diagnosemöglichkeiten und Lernszenarien in Deutschunterricht und Deutschdidaktik* (S. 260-275). Baltmannsweiler: Schneider Verlag Hohengehren.

Neumann, A. (2007). *Briefe schreiben in Klasse 9 und 11*. Münster: Waxmann.

Notenboom, A. & Reitsma, P. (2003). Investigating the dimensions of spelling ability. *Educational and Psychological Measurement*, *63,* 1039–1059.

Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.

Oelkers, J. & Reusser, K. (2008). *Qualität entwickeln – Standards sichern – mit Differenz umgehen.* Bildungsforschung Band 27. Bonn: Bundesministerium für Bildung und Forschung (BMBF).

Oller, J. W. (1976). Evidence for a general language proficiency factor: An expectancy grammar. *Die Neueren Sprachen*, *75*, 165–174.

Ones, D. S. & Viswesvaran, C. (1996). Bandwidth–fidelity dilemma in personality measurement for personnel selection. *Journal of Organizational Behavior, 17*, 609–626.

Oomen-Welke, I. (2003). Entwicklung sprachlichen Wissens und Bewusstseins im mehrsprachigen Kontext. In U. Bredel, H. Günther, P. Klotz, J. Ossner & G. Siebert-Ott (Hrsg.), *Didaktik der deutschen Sprache. Ein Handbuch* (1. Teilband, S. 452-463). Paderborn: Schöningh.

Oomen-Welke, I. (2008). Sprachstandsdiagnose im Elementarbereich: Beobachten, messen und deuten als integrativer Teil der Sprachförderung. In B. Ahrenholz (Hrsg.), *Deutsch als Zweitsprache – Voraussetzungen und Konzepte für die Förderung von Kindern und Jugendlichen mit Migrationshintergrund* (S. 43-64). Freiburg: Fillibach.

Oomen-Welke, I. & Kühn, P. (2008). Sprache und Sprachgebrauch untersuchen. In A. Bremerich-Vos, D. Granzer & O. Köller (Hrsg.), *Bildungsstandards für die Grundschule: Deutsch konkret* (S.139-184). Berlin: Cornelsen.

Open Peer Commentary (2007). Discussion on 'The g-Factor of International Cognitive Ability Comparisons: The Homogeneity of Results in PISA, TIMSS, PIRLS and IQ-Tests Across Nations' by Heiner Rindermann. *European Journal of Personality*, *21*, 707-765.

Ossner, J. (2006a). Kompetenzen und Kompetenzmodelle. *Didaktik Deutsch*, *21,* 5-19.

Ossner, J. (2006b). *Zum Thema des 16. Symposion Deutschdidaktik. Kompetenzen im Deutschunterricht. 17.-20. September 2006* (S. 8-9). Pädagogische Hochschule Weingarten.

Pabst-Weinschenk, M. (2005). *Freies Sprechen in der Grundschule*. Berlin: Cornelsen Scriptor.

Peek, R. (2008). Kompetenzen und Kompetenzmessung im Kontext von Fachdidaktik, Psychometrie und Unterrichtsentwicklung. In Symposium Deutschdidaktik e.V. (Hrsg.), *Sonderheft zum 16. Symposium Deutschdidaktik Kompetenzen im Deutschunterricht* [Sonderheft]. *Didaktik Deutsch, 14,* 162-172.

Peyer, A. (2010). Texttheorie. In M. Kämper-van den Boogaart & K. H. Spinner (Hrsg.), *Lese- und Literaturunterricht* (Bd. 11/1, S. 238-268). Baltmannsweiler: Schneider Verlag Hohengehren.

Pietsch, M., Böhme, K., Robitzsch, A. & Stubbe, T. C. (2009). Das Stufenmodell zur Lesekompetenz der länderübergreifenden Bildungsstandards im Vergleich zu IGLU 2006. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 393-416). Weinheim: Beltz.

Pollock, J. L. & Cruz, J. (1999). *Contemporary Theories of Knowledge*. Lanham, Maryland: Rowman and Littlefield.

Porsch, R. (2010). *Schreibkompetenzvermittlung im Englischunterricht in der Sekundarstufe I*. Münster: Waxmann.

Prenzel, M. (2006). Bildungsforschung zwischen pädagogischer Psychologie und Erziehungswissenschaft. In H. Merkens (Hrsg.), *Erziehungswissenschaft und Bildungsforschung* (S. 69-79). Wiesbaden: VS Verlag für Sozialwissenschaften.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Reich, H. H. (2003). Sprachstandsanalyse und Bildungsreform. In I. Gogolin (Hrsg.), *Pluralismus unausweichlich?* (S. 145-155). Münster, Westfalen: Waxmann.

Rindermann, H. (2006).Was messen internationale Schulleistungsstudien? Schulleistungen, Schülerfähigkeiten, kognitive Fähigkeiten, Wissen oder allgemeine Intelligenz? *Psychologische Rundschau*, *57*, 69-86.

Rindermann, H. (2007). The g-factor of international cognitive ability comparisons: The homogeneity of results in PISA, TIMSS, PIRLS and IQ-tests across nations. *European Journal of Personality, 21,* 667-706.

Robitzsch, A. (2009). Methodische Herausforderungen bei der Kalibrierung von Leistungstests. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 42-106). Weinheim: Beltz.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* (2. Aufl.). Bern: Huber.

Rost, D. H. & Buch, S. R. (2010). Hochbegabung. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (S. 257-273). Weinheim: Beltz.

Roth, H. (1969). Die Bedeutung der empirischen Forschung für die Pädagogik. In S. Oppolzer (Hrsg.), *Denkformen und Forschungsmethoden der Erziehungswissenschaft* (Bd. 2: Empirische Forschungsmethoden, S. 15-63). München: Ehrenwirth.

Roth, H. (1971). *Pädagogische Anthropologie*. Band II: Entwicklung und Erziehung. Hannover: Schroedel.

Ruhloff, J. (2007). Allgemeine Erziehungswissenschaft versus Empirische Schulforschung? Ein Statement zur Eröffnung einer Debatte. In M. Kraul (Hrsg.), *Bildungsforschung und Bildungsreform* [Beiheft]. *Die Deutsche Schule*, *99*, 137-140.

Sailer, M. (2007). Bildungswissenschaft und Bildungsforschung: Eine Rückbesinnung auf den Gegenstand Bildung . In F.-M. Konrad & M. Sailer (Hrsg.), *Homo educabilis* (S. 127-141). Münster: Waxmann.

Sang, F., Schmitz, B., Vollmer, H. J., Baumert, J. & Roeder, P. M. (1986). Models of second language competence: a structural equation approach. *Language Testing, 3,* 54-79.

Sawaki, Y., Stricker, L. J. & Oranje, A. H. (2009). Factor structure of the TOEFEL Internet-based test. *Language Testing*, *26,* 5-30.

Scherner, M. (2003). Grammatik und Textualität. In U. Bredel, H. Günther, P. Klotz, J. Ossner & G. Siebert-Ott (Hrsg.), *Didaktik der deutschen Sprache. Ein Handbuch* (1. Teilband, S. 476-486). Paderborn: Schöningh.

Schiefele, U. (1996). *Motivation und Lernen mit Texten*. Göttingen: Hogrefe.

Schmied-Kowarzik, W. (1993). *Bildung, Emanzipation und Sittlichkeit. Philosophische und pädagogische Klärungsversuche*. Weinheim: Deutscher Studien Verlag.

Schneewind, J. (2007). *Wie Lehrkräfte mit Ergebnisrückmeldungen aus Schulleistungsstudien umgehen. Ergebnisse aus Befragungen von Berliner Grundschullehrerinnen* (Dissertation). Berlin: Freie Universität Berlin, Fachbereich Erziehungswissenschaften und Psychologie. Zugriff am 20.10.2011 unter http://www.diss.fu-berlin.de/diss/receive/FUDISS_thesis_000000002819

Schnotz, W. (1994). *Aufbau von Wissensstrukturen. Untersuchungen zur Kohärenzbildung bei Wissenserwerb mit Texten*. Weinheim: Beltz.

Schrader, F.-W. & Helmke, A. (2004). Von der Evaluation zur Innovation? Die Rezeptionsstudie WALZER: Ergebnisse der Lehrerbefragung. *Empirische Pädagogik, 18,* 140-161.

Schroeders, U. & Wilhelm, O. (in press). Equivalence of reading and listening comprehension across test media. *Educational and Psychological Measurement*.

Schurr, J. (1975). Über den wesensnotwendigen Zusammenhang von Sein und Sollen bei der Bestimmung des Menschen. *Pädagogische Rundschau*, *33,* 3-15.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (1989). *Einheitliche Prüfungsanforderungen in der Abiturprüfung Deutsch. Beschluss der Kultusministerkonferenz vom 01.12.1989 i.d.F. vom 24.05.2002*. Zugriff am 2.09.2011 unter http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/1989/1989_12_01-EPA-Deutsch.pdf

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2004). *Bildungsstandards im Fach Deutsch für den Mittleren Schulabschluss – Beschluss vom 04.12.2003*. München: Wolters Kluwer.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2005a). *Bildungsstandards im Fach Deutsch für den Primarbereich (Jahrgangsstufe 4) – Beschluss vom 15.10.2004*. München: Wolters Kluwer.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2005b). *Bildungsstandards im Fach Deutsch für den Hauptschulabschluss (Jahrgangsstufe 9) – Beschluss vom 15.10.2004*. München: Wolters Kluwer.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2006). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. München: Wolters Kluwer.

Shin, S.-K. (2005). Did they take the same test? Examinee language proficiency and the structure of language tests. *Language Testing, 22*, 31-57.

Shirley, D. (2008). The coming of post-standardization in education: what role for the german Didaktik tradition?. In M. Meyer, M. Prenzel & S. Hellekamps (Hrsg.), *Perspektiven der Didaktik*. [Sonderheft 9/2008]. *Zeitschrift für Erziehungswissenschaft, 11*, 35-45.

Shohamy, E. (1996). Competence and performance in language testing. In G. Bronn, K. Mamkjer & J. Williams (Eds.), *Performance and competence in second language acquisition* (pp. 136-151). Cambridge: Cambridge University Press.

Song, M.-Y. (2008). Do divisible subskills exist in second language (L2) comprehension? A structural equation modeling approach. *Language Testing*, *25*, 435-464.

Speck-Hamdan, A. (2007). Entwicklung von Unterrichtsqualität durch Standards? Überlegungen am Beispiel des Fachbereichs Deutsch in der Grundschule. In K. Möller, P. Hanke, C. Beinbrech, A. K. Hein, T. Kleickmann & R. Schages (Hrsg.), *Qualität von Grundschulunterricht* (S. 91-94). Wiesbaden: Verlag für Sozialwissenschaften.

Spinner, K. H. (2003). Lesekompetenz nach PISA und Literaturunterricht. In U. Abraham, A. Bremerich-Vos, V. Frederking & P. Wieler (Hrsg.), *Deutschdidaktik und Deutschunterricht nach PISA* (S. 238-248). Freiburg im Breisgau: Fillibach.

Spinner, K. H. (2008). Bildungsstandards und Literaturunterricht. In M. Meyer, M. Prenzel & S. Hellekamps (Hrsg.), *Perspektiven der Didaktik*. [Sonderheft 9/2008]. *Zeitschrift für Erziehungswissenschaft*, *11*, 313-323.

Spolsky, B. (1989). Communicative competence, language proficiency, and beyond. *Applied Linguistics*, *10*, 138-156.

Stanat, P. (2002). Spicken erwünscht. *MaxPlanckForschung, 3,* 18-21.

Stecher, B. M. & Barron, S. I. (1999). *Quadrennial milepost accountability testing in Kentucky* (CSE Technical Report *No. 505*). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Steinbrenner, M. (2007). Sprache denken. Eine Kritik an Jakob Ossners Kompetenzmodell. Didaktik Deutsch, 23, S.5-14.

Switalla, B. (2002). PISA lesen. Implikationen der Lesekompetenz-Studie. *Universitas online.* Zugriff am 20.10.2011 unter http://www.heidelberger-lese-zeiten-verlag.de/archiv/online-archiv/switalla.pdf

Tent, L. & Stelzl, I. (1993). *Pädagogisch-psychologische Diagnostik* (Bd. 1., Theoretische und methodische Grundlagen). Göttingen: Hogrefe.

Terhart, E. (2003). Reform der Lehrerbildung: Chancen und Risiken. In I. Gogolin & R. Tippelt (Hrsg.), *Innovationen durch Bildung. Beiträge zum 18. Kongress der deutschen Gesellschaft für Erziehungswissenschaft* (S. 163-180). Opladen: Leske und Budrich.

Tillmann, K.-J., Rauschenbach, T., Tippelt, R. & Weishaupt, H. (Hrsg.). (2008). *Datenreport Erziehungswissenschaft 2008*. Opladen: Budrich.

Tippelt, R. & Schmidt, B. (Hrsg.). (2010a). *Handbuch der Bildungsforschung* (3., durchgesehene Aufl.). Wiesbaden: Verlag für Sozialwissenschaften.

Tippelt, R. & Schmidt, B. (2010b). Einleitung der Herausgeber. In R. Tippelt & B. Schmidt (Hrsg.), *Handbuch der Bildungsforschung* (3., durchgesehene Aufl., S. 9-19). Wiesbaden: Verlag für Sozialwissenschaften.

Vogt, R. (2009). Gesprächskompetenz – Vorschlag eines gesprächsanalytisch fundierten Konzepts. In M. Krelle & C. Spiegel (Hrsg.), *Sprechen und Kommunizieren. Entwicklungsperspektiven, Diagnosemöglichkeiten und Lernszenarien in Deutschunterricht und Deutschdidaktik* (S. 15-40). Baltmannsweiler: Schneider Verlag Hohengehren.

Vollmer, H. J. & Sang, F. (1983). Competing hypotheses about second language ability: a plea for caution. In J. W. Oller (Ed.), *Issues in language testing* (pp. 29-79). Rowley, MA: Newbury House Publishers.

Vonken, M. (2005). *Handlung und Kompetenz. Theoretische Perspektiven für die Erwachsenen- und Berufspädagogik*. Wiesbaden: Verlag für Sozialwissenschaften.

Watermann, R., Stanat, P., Kunter, M., Klieme, E. & Baumert, J. (2003). Schulrückmeldungen im Rahmen von Schulleistungsuntersuchungen: Das Disseminationskonzept von PISA-2000. *Zeitschrift für Pädagogik, 49,* 92-111.

Weiler, H. N. (2002). Die selbstzufriedene Disziplin. Politik- und praxisfern: Die Deutsche Erziehungswissenschaft fördert ideologische Debatten statt zu beraten. *Süddeutsche Zeitung,* (*Nr. 232 vom 8.10.2002*), 17.

Weiler, H. N. (2003). Bildungsforschung und Bildungsreform – Von den Defiziten der deutschen Erziehungswissenschaft. In I. Gogolin & R. Tippelt (Hrsg.), *Innovationen durch Bildung. Beiträge zum 18. Kongress der deutschen Gesellschaft für Erziehungswissenschaft* (S. 181-203). Opladen: Leske und Budrich.

Weinert, F. E. (2001a). Vergleichende Leistungsmessung in Schulen - eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Hrsg.), *Leistungsmessung in Schulen* (S. 17-31). Weinheim: Beltz-Verlag.

Weinert, F. E. (2001b). Concept of competence: a conceptual clarification. In D. S. Rychen & L. H. Salganik (Eds.), *Defining and selecting key competencies* (pp. 45-66). Seattle: Hogrefe & Huber.

White, R. W. (1959). Motivation reconsidered: The concept of competence. *Psychological review, 66*, 297-333.

Wigger, L. (2004). Bildungstheorie und Bildungsforschung in der Gegenwart. Versuch einer Lagebeschreibung. Heinz-Elmar Tenorth zum 60.Geburtstag. *Vierteljahrsschrift für wissenschaftliche Pädagogik*, *80*, 478-493.

Winkelmann, H. & Böhme, K. (2009). Anlage und Durchführung der Pilotierung der Bildungsstandards. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 31-41). Weinheim: Beltz.

Winkler, I. (2005). Zur Beziehung von Unterrichtsmaterial, -gestaltung und -erfolg: Drei Aufgaben zu Georg Brittings „Brudermord im Altwasser" im Praxistest. In J. Stückrath & R. Strobel (Hrsg.), *Deutschunterricht empirisch*. (S. 177-196). Baltmannsweiler: Schneider Verlag Hohengehren.

Wissenschaftsrat (2001). *Empfehlungen zur künftigen Struktur der Lehrerbildung.* Zugriff am 18.08.2011 unter http://www.wissenschaftsrat.de/download/archiv/5065-01.pdf

Wygotski, L. S. (1987). *Ausgewählte Schriften.* Band 2. Arbeiten zur psychischen Entwicklung der Persönlichkeit. Köln: Pahl-Rugenstein.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the Three-Parameter Logistic Model. *Applied Psychological Measurement, 8,* 125-145

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187-213.

Zabka, T. (2010). Texte über Texte als Formate schriftlicher Leistungsprüfung: Nacherzählung, Inhaltsangabe, Analyse, Interpretation und benachbarte Aufgaben. In M. Kämper-van den Boogaart & K. H. Spinner (Hrsg.), *Lese- und Literraturunterricht* (Bd. 11/3, S. 60-88). Baltmannsweiler: Schneider Verlag Hohengehren.

Zedler. P. & Döbert, H. (2010). Erziehungswissenschaftliche Bildungsforschung. In R. Tippelt & B. Schmidt (Hrsg.), *Handbuch der Bildungsforschung* (3., durchgesehene Aufl., S. 23-45). Wiesbaden: Verlag für Sozialwissenschaften.

Zeitlinger, E. (2009). Kompetenzorientierung, Kompetenzaufbau und Nachhaltigkeit. Die Bedeutung von Bildungsstandards für den Deutschunterricht. *ide, 33 (3)*, 51-62.

Ziener, G. (2008). *Bildungsstandards in der Praxis. Kompetenzorientiert unterrichten.* Seelze-Velber: Klett/Kallmeyer.

Zwick, R. (1992). Statistical and psychometric issues in the measurement of educational achievement trends: Examples from the National Assessment of Educational Progress. *Journal of Educational and Behavioral Statistics*, *17*, 205-218.